Meehl, P. E., & Golden, R. (1982). Taxometric methods. Chapter 5 in P. Kendall & J. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127-181). New York: Wiley.

Taxometric Methods

PAUL E. MEEHL AND ROBERT R. GOLDEN

THE THREE KINDS OF TAXOMETRIC TASKS

"Taxometrics" may be roughly defined for present purposes as the branch of applied mathematics that treats of problems of classification. "Classification" in the broad sense includes both the process of constructing or inferring classes ("classification" in the narrow sense) and the sorting of individuals into such classes ("assignment," "identification," "diagnosis"). It is not of course confined to psychopathology or to the classification of persons, or even the classification of living organisms, whether by behavior or other attributes, since stars and stones are also the subject matter of numerical taxonomy (as one can learn by attending meetings of the Classification Society). Like any applied mathematics, the discipline includes metatheoretical concepts—not just a formalism, but an explanatory (interpretative and motivating) text.

The word "taxon" cannot be precisely defined without arbitrariness, as will become clear in the discussion that follows. But as in other fields of science—a point often not understood by social scientists—explicit and purportedly rigorous verbal definitions of a domain or a method are not necessary. The meaning of both theoretical and metatheoretical terms is best set forth contextually, that is, by the formalism and the interpretive text itself. As a first rough meaning-stipulation adequate to delimit our task in this chapter, we may say merely that a *taxon* in psychopathology is an entity, type, syndrome, species, disease, or more generally, a *nonarbitrary class.* The distinction aimed at by this crude definition is that between a taxonic situation and the mathematics appropriate to it, and a nontaxonic situation, such as a dimension. We think of schizoidia or Huntington's Disease as a taxon. We do not think of garden variety social introversion (such as measured by MMPI scale S_i) as a taxon or type, but rather as a dimension.

All taxa are classes, but not all classes are taxa. There are as many classes of individuals in psychopathology, or within normal populations of human beings, as there are cutting scores on dimensions, or conjunctions, disjunctions, and other logical functions of attributes. The purpose of taxometrics is to help the investigator identify and sort those categories of individuals that are in some sense "really in nature," that would be there whether or not clinical psychologists had bothered to take notice of them or were clever enough to detect them. In the famous phrase attributed to Plato, the aim of the taxometrician is to "carve nature at its joints." These crude meaning stipulations and metaphors will suffice for now, because the nature of taxonicity will be explicated in the rest of the chapter.

Initially, in contemplating a domain like psychopathology where it occurs to an investigator to apply some statistical procedure to observations about the

The senior author is indebted to the J. McKeen Cattell Foundation for support of research. We wish to express gratitude to Dr. Larry Krasnoff for his suggestions regarding the description of the taxometric models.

persons' behavior, dispositions, or inferred traits, to sort them into types or clusters or taxa (from here on we shall simply say "taxa"), we can notice three distinguishable states of antecedent information, states of the investigator's knowledge that will appropriately affect his or her strategy and tactics, and, in particular, help in selecting the mathematical formalism appropriate to a taxometric task. We do not here prejudge whether these three are basically different, that is, different *epistemologically, mathematically, causally,* or *pragmatically* (this latter in the context of decision making, what to do with an individual person taxometrically classified by the method chosen). Granting that these prior knowledge conditions may or may not be fundamentally (qualitatively) different, nothing hinges upon that question in our treatment. Perhaps the best way to see whether they are different in important ways is to consider that question *after* having concocted a research strategy associated with a substantive theory, or at the least a substantive metatheory, and *then* to distinguish the conditions contextually or implicitly. Putting it strongly, the character of the distinction between these three antecedent states of knowledge becomes clear only after the research task is well under way. But at the extremes, despite borderline region cases, these states of knowledge seem usefully distinguishable as follows:

I. *Accepted criterion prediction.* Here the investigator knows with high confidence that there is a taxon, and he or she knows how to identify the individual persons who belong to it and who fall outside it. Here "accepted criterion" may designate something definitional, as a clinician literally *defining* schizophrenia as a psychosis involving a specified kind of thought disorder, so that the validity of the sign is, so to speak, stipulative. Alternatively, the investigator may have in mind a sufficiently good empirical causal connection, whether based upon a strong theory or previous research, so that he or she possesses a two-way pathognomonic sign, a trait or symptom serving almost perfectly both as an inclusion and an exclusion test. He or she does not quite want to call this sign definitional; but one can for most purposes treat it as almost so.

   An investigator's willingness to consider a sign stipulative or definitional depends partly upon his or her long-term research program, and partly upon philosophy of science considerations that will be discussed below. Thus, in the case of schizophrenia, for some scholars it *is* a psychotic thought disorder of such and such quality and intensity. For others, thought disorder is a high weight indicator (for us it would be a nearly infallible inclusion test), but it is not definitional because a theory of schizotypy as a personality organization or as involving the specific etiology of a dominant schizogene would provide the explicit definition. Therefore the thought disorder, while a highly privileged indicator, is nevertheless not definitional for scholars holding those views. We should emphasize also that most philosophers of science today would question whether a clear distinction can be made here, in the light of current views concerning the implicit or contextual definition of theoretical entities in empirical science.

II. *Classical cluster analysis.* In this situation, the kind typically envisioned by investigators in the Classification Society, there is no known taxon having

an accepted criterion as in Case I, nor is there a conjectured taxon whose existence and indicators the investigator considers well enough corroborated so that he or she will allow it to play a major role in the taxometric search procedure. Instead the investigator has a kind of "metaconjecture" to the effect that taxa (usually plural in this knowledge state) do exist in the domain of entities under study, and hopes that he or she has succeeded in concocting a set of domain-relevant indicators for his or her initial list. In the extreme case of no theory or even theory-sketch, the provisional indicator list may have been made deliberately to include every attribute of the entities in the set on which there is any appreciable variation among individuals. If this has been done, and no initial differential weights are assigned on theoretical or clinical grounds, we have the extreme case of what some have complained of as "blind" numerical taxonomy. Such a list of unweighted indicators, in which we ask how many different ways one honeybee may observably differ from another, is that advocated by such eminent contributors as Sneath and Sokal (1973).

III. *Conjectured latent taxon.* Here the investigator has some degree of purported knowledge from his or her clinical experience, or statistical study of file data, and maybe the sketch of a theory, which leads him or her to conjecture the existence of a taxon. But the taxonic conjecture is highly problematic, and furthermore the investigator may have only moderate confidence in his or her list of admittedly fallible indicators, since their validity is conjectural even if one were to grant for the sake of argument the reality of the taxon itself. It is improbable that anyone would conjecture the existence of a latent taxon unless he or she had noticed clinically or detected statistically some kind of co-variation of behaviors.

In Case I, accepted criterion of the taxon's membership, what is the motivation for a taxometric analysis? Why trouble ourselves to construct a linear discriminant function or some nonlinear function to discriminate the groups, or a Bayes' Theorem inverse probability from symptom pattern, or a function-free actuarial table such as advocated by Lykken (Lykken, 1956; Lykken & Rose, 1963)? It is initially puzzling, since we possess an accepted criterion, either definitionally or from previous causal knowledge, leading to an extremely high indicator weight so that there exists an indicator that is almost infallible, why we would concoct a function or table of fallible indicators. We already know that the taxon exists, and we already know "who is who," that is, which individuals belong to it and which fall outside it.

One reason we might want to do this is in cases of *concurrent validity* (Cronbach & Meehl, 1955), in which the accepted criterion used in identifying individuals as being within the taxon or outside it is too costly, painful, or dangerous to collect routinely on all patients or on all persons being screened, but in the investigator's research context this (normally unavailable) accepted criterion is available on all cases.

Another reason is in the *predictive validity* situation, in which the accepted criterion that defines taxon membership becomes available only at a later point in time. Ascertaining its presence is not costly, painful, or dangerous, but we may

have good clinical or other reasons to try to make a decision in advance as to that person's taxon membership prior to the time when the defining sign appears. An example is forecasting who will develop Huntington's Disease among a group of individuals known to be at risk because of having a Huntington's afflicted parent, when the subject of genetic counseling must decide whether to marry and have children (see Lyle & Gottesman, 1977). Again, if one anticipated an epidemic of a disease in which only a minority of patients develop a certain dangerous or residual-producing complication, immunization against the disease might be strongly advised for persons especially prone to the complication. This sort of example is one that influenced Carnap in his classic paper on testability and meaning (Carnap, 1936, 1937) to distinguish *confirmability* and *testability* as positivist meaning criteria, and to advocate confirmability as the more "tolerant" criterion. If I can't produce the disease, I cannot (at will) *test* the dispositional hypothesis "Jones (presently not ill) is prone to complication C should he fall ill with disease D"; but I may yet *confirm* this hypothesis, because *if* he should fall ill with the disease, then we will discover whether he develops the complication or not. We do not wish to deny empirical significance to the statement "Jones is C-prone" prior to his falling ill, partly for clinical (prophylactic) reasons, but also because it is by attaching significance to the confirmable but untestable C-statement that we motivate a research strategy that includes search for antecedently available indicators of his complication proneness.

A third possibility, more common in economics and not there usually taxonic, but at least imaginable in psychopathology, is the case in which the qualitative nature of some of the fallible indicator variables put into the taxometric equation or actuarial table makes them plausible candidates to be *causative,* so that one may attach a causal weight or influence interpretation to the beta coefficients. An important variant of this one in psychopathology, as in schizophrenia research, is the possibility that one of the indicators of high weight, while not causative itself, is a pleiotropic indicator of the causal schizogene, for example, a soft neurological sign, or the alleged anomaly in nail-fold capillary bed in the nonpsychotic relatives of schizophrenic probands (Buchanan & Jones, 1969).

Why classify people at all? Arguably, part of the problem in taxometrics, such as the low yield of formal cluster algorithms in psychopathology research to date, lies in failure to put this initial rock bottom question before proceeding to invent interperson similarity measures and search procedures for clustering within such similarity matrices. One easy case to defend is that of purely administrative convenience. If we have a cutting score on an engineering aptitude test, we may sort people into "good risks," "probationary risks," and "poor risks—reject applicant." But nobody conjectures that these are taxonic, either in terms of the predictor variable or in terms of the ultimate criterion, since an administrative change in the honor point ratio required to remain in engineering school would reclassify individuals on the borderline. Despite its unavoidability, especially in legal contexts in which what the psychologist sees as dimensions are usually replaced by quality words (predicates, categories instead of continua), this situation has no theoretical importance and will not be discussed further.

A second consideration is convenience in communication, as when we

summarize a great deal of our knowledge, including low probability or highly problematic conjectural knowledge about a patient, by saying "she is hysteroid" or "he is a borderline sociopath" or "so-and-so is a latent manic-depressive." It is doubtful that clinicians could talk with each other if they were strictly forbidden to employ such summarizing rubrics as this, although the dangers of reification when the entity has no existence, as well as premature diagnostic closure by one's adoption of the semantics, are well known.

Another case is somewhat more substantive, although related to the communication or summarizing function; that is, that *we don't have to gather all the research data on patients in one huge project.* Example: No one seriously imagines that there are fewer than, say, 100 test scores, traits, or life history facts that would be needed to characterize an individual adequately for most clinical purposes, including features such as the disposition to respond favorably to one drug rather than another, suicide risk, therapy stayability, job satisfaction, and the like. But even such a carefully pruned list of attributes means potentially 4950 pairwise correlations among them. Very few clinical research projects have sufficient staff and money to process a sizeable number of patients with respect to 100 variables. It would be wasteful if we lacked some way of allowing investigator X in Albuquerque to relate two drug responses to the amount of schizoid component in manic-depressive patients, and later make use of the fact, when we survey the research literature, that investigator Y in Minneapolis has discovered two new soft neurological signs of the schizoid component, although Y has not studied the therapeutic effects of psychotropic drugs on his cases. It is obvious that organic medicine would have progressed much more slowly had it not been possible for different clinicians (and pathologists, biochemists, epidemiologists, geneticists, physiologists, and bacteriologists) in different places, studying nonoverlapping groups of patients, to mediate the connection of their investigations via a diagnostic rubric. In terms of the difference between input and output variables, instead of having mn input-output correlations to investigate, we may substitute the stochastic relationship between the m indicators and the nosological entity on the input side, and the relationship between the nosological entity and the n indicators on the output side. There are then $(m + n)$ relations to be studied instead of (mn) of them, a considerable economy in both communication and research costs.

Another consideration is a clumping in two kinds of spaces, say, as the space of job descriptions and the space of personal traits. Large-scale personnel contexts such as the military have found it useful to group jobs on the basis of a statistical study of the components in a job analysis, attempting to map, even though only probabilistically, clusters of enlisted men (in terms of their tested interests and abilities) against the job clusters. There is here an economy of transfer and retraining, and we note that might be the case even if there were no marked degree of "clumping" in the two phenotypic hyperspaces, although the more clumping there is, the more saving is achieved. But even if there were homogeneity in the space, it might still be worthwhile to do it.

Finally, and of most interest to the present authors in connection with schizophrenia research, there is the purpose of theoretical (= *causal*) understanding. A side aspect of this is file research, in which a statistical power function

on some variable of interest may be greatly improved by homogenizing the patients, although failure to homogenize the patients would not constitute a vitiating error in most studies, contrary to what some clinicians have alleged. Nevertheless, if we are looking for a weak effect, or even an effect that is reversed in some categories, the fact that the patient is not available for assessment makes it desirable to find nosological discriminations recorded in the chart.

When we speak of a taxon as "latent," what meanings might this have? One meaning is environmental stimulation that has not yet been made explicit or discovered. This would be the only kind of latent taxon readily allowed by strong Skinnerians. Another meaning of "latent" in the psychopathology context is historical, a slight stretching of the meaning, but indicating that the factor is not presently visible in the patient's behavior or surround. It is an event in the past that may or may not be in the record, and in fact, even in extensive uncovering therapy as classical analysis, may or may not come to light as an alleged fact of memory. Examples are a battle-axe mother; Freud's early theory of prepubertal sexual seduction as the specific etiology of hysteria and the obsessional neurosis (which disorder, of the two, depends upon whether the future patient was mainly passive and experienced fear or disgust or mainly aggressive and experienced erotic pleasure); head injury; undiagnosed mumps encephalitis; and the like. These examples show that a latent taxon can be connected with a specific etiology that does not have to be a gene, or a germ, or a "disease" in any strong sense of that word. If Freud's theory had been factually correct, the specific life history event could have been taken as the stipulative definer of hysteria, or at least given such a high weight that a patient might be refused the diagnosis, even if he or she presented conversion phenomena when seen.

Another meaning of "latent" is a causal factor within the person. This is perhaps the most natural meaning for the clinical psychologist. By "within" is meant literally *within,* that is to say a gene, a germ, a brain tumor, a psychological complex, an unconscious fantasy, a repressed memory, and the like.

CLINICAL DISCOVERY OF SYNDROMES

With these possible meanings in mind, let us reflect metatheoretically on a hypothetical example outside of psychopathology, namely, the discovery of a new organic disease. Dr. Fisbee is a practitioner with research interests (for a nice illustration in psychopathology, see Freud's classic paper on the anxiety neurosis and his reply to Loewenfeld's criticism of it, Freud, 1895a,b). Dr. Fisbee is struck with the fact that in the last few months he has seen five patients who presented with complaints of headache and spots before the eyes, and on examination had a low-grade fever, a slightly purplish tongue, and pink ears. He hadn't really "noticed" this pattern until seeing the fourth patient, and the fifth patient seemed to corroborate that he was seeing something orderly that he hadn't studied in medical school. So he goes back to his charts and satisfies himself that he has in fact seen five patients with the combination. He then remembers a few others who had presented only a part of the combination; for instance, in several the purple tongue, the subjective complaints, and the elevated temperature were present, but the pink ears were not. He also reflects that he misdiagnosed some of these

patients as having some other disease. On reading the chart notes for those who came back, as he asked, for follow-up, he finds that the subsequent course didn't fit the diagnosis that he had made of some entity already known to him. So he publishes a clinical note, and Fisbee's Syndrome gets into medical thinking, and then into the literature; and Dr. Fisbee is ultimately immortalized in Dorland's Medical Dictionary for this discovery. Although the particular syndrome is hypothetical, we are not here sketching out an improbable state of affairs, but on the contrary exemplifying the usual course of clinical investigation in organic medicine (see, e.g., Major, 1932).

How does Dr. Fisbee think about it, definitionally and in terms of causal understanding, and more importantly, how about the university's internal medicine professor who, having read the article, decides to research this entity? If one is a hard-line operationist, as some psychologists profess to be, although they are almost never consistently so when pressed, one wants first to arrive at a "clear operational definition" of Fisbee's Syndrome. Presumably this would mean specifying the list of defining symptoms. At this stage of our knowledge, we do not know what is the cause of the syndrome, and we may not even have any conjectures, although the fact that elevated temperature is part of it and that it runs a short self-limited course (nobody has died of it yet), makes us think it might be an infectious disease; but that's not totally clear. In any case, we do not know what kind of germ is involved, and we as yet do not have any tissue pathology. So one might say that the definition of Fisbee's Syndrome lay wholly in the symptom list. That's misleadingly simple, and it's wrong. Seeing why it's wrong will help the reader see why some of the statements made by psychologists about entities like schizophrenia that purport to be sophisticated are actually naive and based upon an undergraduate-level philosophy of science, and, it would seem, woeful ignorance of the history of developments in biology and medicine. We can't simply say that the "list of symptoms" operationally defines the entity. We have to say how that list is put together in the definition. We have symptoms S_1 (headache), S_2 (spots before eyes), S_3 (low-grade fever), S_4 (purple tongue), and S_5 (pink ears). First we think of a simple logical conjunction of the symptoms as defining Fisbee's Syndrome. Will that work? No, it will not. Try to write an explicit operational definition of disease D as $D = S_1 \bullet S_2 \bullet S_3 \bullet S_4 \bullet S_5$. This of course means that any patient lacking any one of the five symptoms is excluded. So the patient Dr. Fisbee remembered on reflection, who presented with the complaints of headache, spots before the eyes, purple tongue, and temperature elevation, but whose ears were not pink, cannot be "counted" as a case of Fisbee's Syndrome. Now admittedly this could be treated as a totally stipulative matter in the fashion of the old-line positivists, who thought that a definition simply "recorded one's decision to use a word in a certain way." (We are not here arguing about words, or about the correct view that definitions are *in some sense* ultimately stipulative.) Why does not the sensible physician or medical researcher consider doing it in this way? It's really very simple. In the history of medicine it has been repeatedly found upon thorough comprehension of a disease—in all of its important aspects, including its pathology and specific etiology, its course, the reasons why certain patients get well and others don't, and why certain therapies work in some patients and not others,

all of them sharing the same disease when it is thoroughly understood—that *there are very few symptoms that are absolutely two-way pathognomonic.*

Our Dr. Fisbee, the research-oriented practitioner, may have no interest in philosophy of science. He may never have heard of Carnap's discussion of reduction sentences, or Pap's classic fundamental paper on open concepts (Pap, 1953; 1958, Chapter 11). But he is a sensible person, not having been brainwashed by some undergraduate social science course that delivered a simplistic view of "operationism," and he knows from common sense considerations that the color of somebody's ears is not likely to have a direct causal influence on the color of his tongue, and that neither one can produce a fever. Therefore, he reasons, if these various externally visible phenomena (together with the subjective complaint of headache and spots before the eyes) have a tendency to go together—both in the sense that they are found in the same patients with a disproportionate frequency, and in the *P*-correlation sense that they covary over time in the same patient as he falls ill and recovers—*then there must be some common causal factor that produces all of them.* Unless that common causal factor is sun spots or witchcraft, which Dr. Fisbee rationally doesn't believe in, where does that common factor have to be? Well, if it's in the world of space and time and material entities, the only place for it to be is inside the patient. That "something or other," whatever it is, is the internal cause of the syndrome. In organic medicine it is the *pathology* of Fisbee's Syndrome. A further conjecture is that this pathology is in turn *caused* by something specific, and that something specific may be in the person (as a mutated gene or a toxin or a germ) or a historical event such as trauma. A combination of these commonsense considerations with his knowledge of the history of medicine leads a rational medical researcher to the conjecture (which he hardly treats as a conjecture but as a near certainty, despite his slight knowledge of the newly discovered entity) that a *specific etiology-cum-pathology* underlies the phenotypic syndrome that first called his attention to the disease. To such a sophisticated research-oriented doctor, the cliché remark of some psychologists that it is "tautologous" to refer a symptom to a disease entity called schizophrenia is not sophisticated but merely ignorant.

So the reason that a conjunction of symptoms won't do is that it's *too strong*. That is, it would result in too many false negatives, since every symptom must, on that view, be an infallible exclusion test. But the history of medicine, as well as our overall knowledge of how the innards of the body work makes it immediately comprehensible why the relations between the symptoms pairwise are only stochastic. The reason that the pairwise correlations of the five symptoms are stochastic, although some of them may be high (if the syndrome is "loose," it will be harder for Dr. Fisbee to notice it), is that they all have a merely stochastic connection with the specific pathology, and that fact results in their imperfect correlation among themselves.

Seeing that a conjunction of symptoms as the explicit operational definition of disease *D* is too strong, it then occurs to a philosophically oriented pathologist to substitute a *disjunction* of symptoms, that is, we try $D = S_1 \vee S_2 \vee S_3 \vee S_4 \vee S_5$. But a disjunction is just as bad as a conjunction, and in fact somewhat worse, because whereas a conjunction was somewhat too strong, a disjunction is grossly too weak.

It suffers a vast excess of false positives, because it treats each of the separate elements of the syndrome as if it were a perfectly valid inclusion test. Notice how social science is unlike medicine in this respect. Most social scientists wouldn't like to pay much attention to a sign that is often present in a great variety of diseases. But no physician wants to drop a symptom like elevated temperature out of the list of symptoms of, say, an infectious disease, despite the fact that there are literally hundreds of conditions—and not all of them infectious diseases—that may produce fever. We don't want to leave elevated temperature out of our set of indicators for measles or malignancy or uremia, but obviously a disconjunctive definition of any of the diseases with elevated temperature as one of the disjuncts would result in a ratio of false positives to valid positives that would be an order of magnitude or two greater!

One important difference between the way organic medicine makes inferences and the way the psychologist or sociologist typically does is that a logical tree, with stepwise exclusion of possibilities, means that a sign may not be attended to at one stage in the diagnostic process but becomes critical at another stage because, while not powerful as an inclusion test, it is a powerful exclusion test. This is a different way of proceeding from an assignment of beta weights, or the unit weighting of low validity test items in a personality inventory. An inclusion test is a sign that is almost never present unless the disease is present and whose manifestation therefore permits us to infer the disease with high confidence. An exclusion test rules out the disease, in that it is almost always present when the disease is present, so its absence makes the disease almost certainly absent. "Pathognomicity" in medicine, when not otherwise defined, more commonly seems to mean an inclusion test; but the strong meaning is two-way pathognomicity, that is, a sign whose presence rules the disease in and whose absence definitively rules it out. Some social scientists assume, quite wrongly, that two-way pathognomicity abounds in organic medicine, but we have yet to find any physician who claims to know of many literally pathognomonic signs for the various organic diseases.

Why is not the entity the same as the syndrome? Passing the inept objection about circularity, the reason that the entity is not the syndrome is that we have a conjectured latent cause, which we confidently anticipate will, given adequate research, some day be the explicit definition of the disease; but that pending that research outcome, the disease is an *open concept.* What we are saying, roughly, is in the form of a scientific "promissory note." We take it for granted that there exists a unitary underlying latent pathology-cum-etiology that gives rise to the correlation of the symptoms (S_1, . . . S_5) but we don't yet know what that latent causal factor is. When we do know it, it will become definitive of the disease. At present, the disease *D* is not strictly defined. It's a fuzzy notion, an open concept (Pap 1953, 1958; Meehl 1972c, references cited p. 21). Its meaning is not empty because it is stipulated as partly *that which underlies the production of Fisbee's Syndrome,* as probabilistically linked to each of the phenotypic facets of the syndrome. But obviously it would be bad semantics to claim that we are *explicitly defining* "Fisbee's Disease" by Fisbee's Syndrome, or even that we're claiming that nobody can "have the syndrome" if any one of the elements is missing

(conjunctive stipulation); or even worse, that anyone can be said to have the syndrome whenever any one of the elements is present (disjunctive stipulation).

There is an interesting and instructive oddity about symptom percentages in a medical book being presented for a disease whose specific pathology and etiology are not yet known. A hyperoperational critic might ask, what does it mean to say that a certain percentage of patients have each symptom of the disease when the entity consists of "nothing but" the symptoms? The answer, of course, is that the disease entity does *not* consist of "nothing but" the symptoms. The disease consists jointly of the (unknown, conjectured) latent pathology-cum-etiology taken together with the symptoms, this "taking together" being stochastically understood. If that Papian notion is too subtle or complicated for a psychologist, he or she should perhaps pursue some other area than psychopathology, since nature apparently is not simple enough for such a simplistic philosophy of science.

So we see that doctors were thinking about open concepts and promissory notes before a philosophy of science that treats of these notions came to be invented, and certainly long before any formal mathematical taxometrics was invented. It is worthwhile, however, to pursue the usual superoperationalist objection a bit further. While we are forced to grant that a conjunction is too strong and a disjunction too weak to make clinical or theoretical sense and to motivate meaningful research into the inner nature of Fisbee's Syndrome, it does seem a little strange to say that a person can have Fisbee's Disease without showing a single element of it. Now, of course, if the patient doesn't present any symptoms or complaints at all, he or she won't come to the doctor (or, if in a routine physical, no clinician will think anything's the matter with him or her). Nevertheless, we do have to make some philosophical place for the accepted idea of a *silent disease*. In organic medicine it is taken for granted that diseases can be symptomatically silent. For instance, consider a person who has never gone to a physician complaining of any of the usual kidney symptoms, let alone the excruciating pain of a renal calculus. At post mortem, following being killed by a truck, it is discovered that the entire kidney lumen on one side is totally occluded by a huge mass of potassium phosphate and carbonates, the so-called "staghorn kidney." In an interesting study of his own diagnostic errors, a high-caliber Minneapolis internist (Peppard, 1949) showed that even being very hard on himself about the cause of his errors, the commonest single source of omission diagnostic error was literally *symptoms or signs not found,* meaning that he had carefully looked for them properly, with the best available techniques, and the patient simply did not show the sign. It is strange that psychologists would think that the mind is so much simpler than the innards in organic medicine that, whereas the idea of a neurologically silent brain tumor or a silent carcinoma of the liver or a silent staghorn kidney can exist, yet it couldn't be possible that there could be a silent schizoid component or a silent psychological complex or a silent temperamental disposition!

If we understand *why* each pairwise correlation is stochastic—because each sign is only probabilistically linked to the (unknown) causal factor that will ultimately become definitive of the entity—it is obvious that a certain probability attaches to each of the possible configurations, including having all five symp-

toms, four but not five, three but not the other two, and so on. Then it is quite arbitrary—given that causal model and our epistemic situation—to cut at having one symptom, at "symptom score" 0/1. Because the probability for each symptom S_i being $p(S_i/D) < 1$, *the joint probability for absence of all five S's is not zero,* excluding a very unlikely configural effect not disease-related. Consequently, a person can have Fisbee's disease ("silently") even if none of the symptoms is present. There is nothing complicated about this, once one understands the distinctions among epistemology, statistics, and causality.

Notice that the existence of degrees of a trait as an indicator of a latent taxon does not preclude taxonicity. One still hears the stupid objection that there can't be an "entity" schizophrenia because all degrees of thought disorder (or inappropriate affect, or withdrawal) exist. This is about as bright as saying that there can't be such a thing as meningitis since one of its exclusion tests is high fever, and all degrees of fever exist in sick people! We hope it is not necessary to treat further of that pseudo-sophisticated complaint.

CLINICAL TAXA AS OPEN CONCEPTS

This is not the place to develop a detailed philosophical analysis of open concepts, but readers are urged to look at the paper of Pap (1953; 1958), and, if the subject interests them sufficiently, the set of related articles cited in Meehl (1972c, p. 21). It is almost universally accepted among logicians and historians of science today that most theoretical concepts are not defined operationally in Bridgman's original sense, but are defined contextually or implicitly, that is, by their role in a network of nomological or statistical "laws." The possible exceptions to this statement are still in debate, and they are so rare, even in highly developed sciences like physics and chemistry, that it seems rather pointless to argue the question for psychopathology. Thus to say that intelligence is what an intelligence test measures, or that the intelligence factor *g* is adequately "defined" (in any strong, strict meaning of that term) by the subtests of an intelligence test, is an oversimplified and misleading account. But in a richer and subtler sense, it *is* contextually "defined," although careful usage would say that the meaning is partially given or stipulated by the theoretical network, however tentative and as yet impoverished that network may be. The "meaning" thus contextually provided is an open concept meaning, a partially specified notion, so its denotation is a "fuzzy set." We hold with Sir Karl Popper (1962, pp. 18-21; 1974, pp. 12-23) that stipulating meanings is a trivial and unimportant exercise compared with testing theories, and that the scientifically valuable components of the former are normally achieved via the latter. Crudely put, you know what you mean by an entity to the extent that you have a set of concept-overlapping statements in the theoretical language, and a proper subset of those statements are linked to statements in the observational language. These statements are, in different ways, all about the entity—where it's found, what brings it about, what it does, what are its various properties (including "compositional properties," what it's made of and how). Only a few of these properties are directly tied to observables. A reader with philosophical interests should have a look at Carnap's classic paper on testability and meaning as a starter into the theory of open concepts, and then read Pap.

Three kinds of openness of concepts, which are not equivalent but which are related, are described in Meehl (1977, 1978). The first kind of openness is the kind expounded by Carnap, and further developed by Pap. It arises from the fact that even a *dispositional* concept (e.g., solubility, dominance, or, in psychopathology, depression proneness) is specified by a *list* of indicators or symptoms (whether testability or confirmability be imposed as a condition of scientific acceptability), and such a list is extensible. The extensibility of the list of indicators was one reason why Carnap introduced the idea of reduction pairs as a substitute for explicit definitions in his original classic paper. We find out that if a wire touching a frog's muscle makes it twitch, a compass needle brought near the wire will be deflected. So we now have two indicators of the latent mysterious "electric current" (long before we knew about electrons, let alone the Fermi mathematical theory of the electron gas). It then turns out that a wire that has these two properties will also, if snipped and immersed in a silver chloride solution, deposit silver at the cathode. If connected with a filament in an exhausted bulb, it will produce light; and so forth. As Carnap points out, although each of the reduction sentences added to the list might be looked upon as, in a sense, a definition (a convention, a meaning specifier), when we put two or more of them together, we are automatically committed to a synthetic claim about a question of fact, that is, that the indicators will agree. That is one of the earliest showings by a logician that the distinction between definitions and empirical statements is not a clean one in the theoretical sciences. In psychopathology we might originally give a tendency to striped muscle conversion reactions a privileged status in identifying the hysteroid disposition. But subsequent clinical experience and psychometric research leads us to consider a patient's preference for certain defense mechanisms (e.g., repression, denial) and various other attributes of the personality structure (impunitive reaction to frustration, a tendency to manipulate the environment by histrionic threats and gestures, a combination of hypersexual signaling with frigidity, etc.) as more privileged indicators than whether one uses a paralyzed arm to get out of washing the dishes.

A second kind of openness applies even to the single indicator, namely, that the connection between the latent entity (or even the open concept, without explicit reference to the promissory note of causality) is probabilistic rather than strict or nomological in character. This state of affairs is even true in the physical sciences, as Pap pointed out in his classic paper on why we can't have a strictly operational closed-concept definition of temperature. Nobody can stick a thermometer into the center of the sun, but we need to talk about the sun's temperature, and we don't wish to say or imply that this is a totally different concept from the temperature of a blast furnace or a bowl of soup. A fortiori, in the life sciences we have to rely on alternative epistemic paths. Any complicated organic system presents occasions for nuisance variables or random factors, usually both, to "get in the way" between the conjectured specific causal entity (germ, gene, complex, drive, memory, or whatever) and our fallible phenotypic indicators of the latter. They have to be fallible, because they are connected with the causal entity by a chain whose links always include stochastic rather than strictly nomological links. This is why one can give strictly operational definitions of most psychological

attributes only by deceiving himself, or alternatively, by confining himself to the most trivial instances and insisting, by God, that the mind had better be simple because that's the way we want to think about it!

A third kind of openness in concepts Meehl has dubbed "Orphan Annie's Eyes." In the usual logician's diagram of the nomological network, the strands of the net are the conjectured laws connecting the conjectured entities between which the laws hold. The entities are the nodes of the network, and are implicitly defined by their role in the network. Those nodes of the net are drawn as little circles, and remind one of the vacuous eyes in the juvenile fascist "Orphan Annie," in a comic strip by that name during the Great Depression. The openness of these "Orphan Annie's Eyes" nodes in the nomological net corresponds to the fact that we don't know the inner nature, that is, the composition and structure, of the theoretical entities insofar as they are only contextually defined by their role in the net.

The gene provides a beautiful instance of "filling in Orphan Annie's Eyes." The gene begins as a Mendel *factor,* as a conjectural entity that is latent with respect to the observed statistics of fertilization in garden peas and the phenotypic traits of fruit flies. The discovery of linkage maps and the realization that the number of linkage groups corresponds to the number of chromosomes in a species makes it natural to conjecture that the "factors" are entities having a physical location at a certain position along the chromosome; hence we have the conjectured *gene.* Further evidence then shows us that cytological anomalies (e.g., translocation) are associated with pronounced aberrations in the linkage statistics of a fly strain. Finally, with the discovery of the giant chromosome in the salivary gland of *Drosophila* we are able to see that X-ray bombardment resulting in a recessive acting strangely like a dominant suggests that there ought to be a "hole" at a certain place on the giant chromosome. Sure enough, direct cytological study shows an "empty place" there in the predicted position!

Notice that every one of these operations, while it locks in the concept of a gene by a very nice and intellectually satisfying nomological network, nevertheless still has the gene as an Orphan Annie's Eye. Only with the solution of the DNA problem by Crick and Watson, where we now *explicitly define* the term "gene" as a cistron, and a cistron as a certain sequence of codons, and a codon as a certain sequence of the four bases, adenine, guanine, cytosine, and thymine, do we fill in Orphan Annie's Eyes. For the first time we can offer a truly explicit definition of the word "gene." It takes us over half a century to get there, and a very powerful and intellectually exciting science of genetics existed prior to Crick and Watson's solution of the double helix. Even that definition is not operational in the usual sense, because the term "gene" is itself defined reductively, that is, in terms of theoretical concepts from another science, namely, organic chemistry. We note, finally, that it is now conceptually detached, so that a gene made in the laboratory by human molecular biologists, and not put into a protoplasmic surround where it is capable of controlling anything phenotypic as an indicator, would still be a gene in the full sense of the word.

It is foolish in psychopathology to pretend that there is any magical way, by either statistics or verbal definitions, to avoid these three kinds of openness. Part of the purpose of a taxometric formalism is to tighten up the open concept by

refining the mathematical characteristics of the net. Putting that together with an embedding explanatory text tells us something about the entities that the abstract notation of the formalism denotes and quantifies.

We may have spent more time on general methodological considerations than necessary, but experience in attempting to explain our own new taxometric methods, and reading of the controversial literature on diagnostic categories (e.g., "labeling theory"), combine to convince us that a certain amount of Augean stable-cleaning is needed as propaedeutic to an intellectually respectable treatment of the taxometric problem in either its mathematical or epistemological aspects.

FORMS AND DEGREES OF TAXONICITY

We now approach the difficult question: what is *taxonicity* as a concept? After several years of philosophical and mathematical effort, plus surveying taxonic entities in several fields other than psychopathology, we have regretfully concluded that the reason it is so hard to give a good definition of "taxon" is that taxonicity itself is not taxonic but is a matter of degree. We use the word "taxonic" rather than "taxonomic" because taxonomy is the *metatheory of classification.* Strictly speaking, it is an abuse of language to speak of a taxonomy as being the taxonic entities themselves. The taxonomy is the taxonomist's science of taxonomizing; the taxa are entities that (if he or she is right in his or her conjectures) exist in the world. A fact-situation is not taxonomic; it is the investigator's approach that is taxonomic. When we refer to the state of nature that the scientist conjectures to exist when proceeding taxonomically in his or her research, that state of nature should be called "taxonic."

All efforts that we have made (and they will not be reported here) or that we have seen others present, whether in psychopathology or in the other life sciences, to give a purely phenotypic characterization of taxonicity are fuzzy at the edges and liable to misinterpretation. This holds true even for those "conventionalist" or "economical" taxonomies that avoid theory and causality and confine taxonic interpretation to densification in the phenotypic hyperspace. Such a clumping or clustering admittedly is what normally leads the clinician or researcher to notice a conjectured taxon in the first place. The initially unpleasant truth of the matter—it turns out to be pleasant, insofar as it relieves us of the burden of giving an explicitly operational definition of "taxonicity" that will satisfy everybody—is that there are several kinds of causal paths to generation of a clumping or clustering of phenotypic characteristics in the descriptive hyperspace, and some of them are not interesting theoretically or valuable pragmatically. Any factor, including arbitrary social selection factors that determine subpopulations to come before our eyes, that can densify the descriptive hyperspace in a certain region more than one might have anticipated by contemplating, say, the Pearson *r*'s of the indicator variables taken pairwise (calculated over the whole realized space) can thereby produce a situation that one might legitimately call "taxonic" in one of the several senses presented above. Example: Most of us probably would think of "being a bridge player" as taxonic; witness the way the question is put: "Do you play bridge?" The person queried might answer, "Yes, but not very often." or "Not very well." But we do not expect him or her to say, "I don't know, it depends on your

cutting score." It is an empirical fact that if a person knows the word "renege," he or she will ($p > .99$) know the words "vulnerable" and "slough." Those who know how to respond to a certain bid will almost certainly have a statistical tendency to avoid certain ways of dealing with the cards in the dummy if they are playing the hand. The taxonicity of some acquired skill patterns is so "tight" that industrial psychologists can make trade tests using as few as eight or 10 verbal items (e.g., "What is a chuck used for?", identifying lathe operators). If one constructed an achievement test of terms from contract bridge playing without deliberately rigging it, employing a suitable kind of random choice from the words appearing in a bridge manual glossary, he would undoubtedly find a clearly bimodal distribution and would conclude that bridge playing was taxonic. It is nevertheless true that no matter how many such items we added in order to slice the pie finer, there would be an admittedly tiny (but not zero) number of scores in the valley between the two curves. They would be people who came to have these intermediate achievement scores in special and unusual ways. For example, a man goes hunting with his friends; they are snowed in; and his three friends insist on teaching him bridge. He doesn't want to learn, but he goes along to be cooperative. He isn't grabbed by the game, and he has never played it since. Such a person could have a score in the valley between the two modes.

A more distressing example of pseudo-taxonicity is what would happen if one foolishly threw together scores on the WAIS subtests for inmates of a state institution for the mentally retarded with a random sample of normals. There would be an appearance of a big taxon, but only some of the people in that low IQ group would represent truly taxonic entities, such as one of the Mendelizing mental deficiencies, or one of the developmental anomalies. Others would simply be the low end of the normal polygenic intelligence curve, appearing as a taxon because of the fact that an institutional population was thrown together with a general population of "normals." There is no point in multiplying examples. The simple fact is that there are various ways in which a phenotypic clumping or clustering in a descriptive hyperspace can be achieved. There are as many different defensible meanings of "taxonic" as there are causal origins of clumps or clusters.

Our own preference, which we have no desire to force upon others, is that if a strong meaning of the word "taxon" is to be adopted, it should be like that of medicine and genetics, namely, a *causal-theoretical meaning.* But even if this convention is accepted, it doesn't solve the problem. One may prefer to think of a true taxon—one that carves nature at its joints and is not merely administrative, or the demarcation of a region on a dimension for communicative ease, or because the courts understand category concepts better than metrical concepts—as causal, involving a *specific etiology.* We then have to ask the further metatheoretical question; what do we mean by the phrase "specific etiology"? Meehl (1977) has set out a series of meanings of "strong influence," only the strongest of which are specific etiologies as that term would be used in, say, medicine or genetics.

In a causal interpretation of strong taxonicity, we define the latent taxon in theoretical terms. That is true of our conceptualization and our research strategy, even if the theoretical terms themselves are only vaguely contextually understood at a given stage of the research enterprise. This means that we need not struggle to

attain a definition of taxonicity by reference solely to the phenotypic distribution, that is, by the distribution properties of the fallible indicators. Furthermore, we know (see, e.g., Murphy, 1964, and the example above) that a latent taxon can exist, as theoretically comprehended with reference to its causality or inner nature, without generating a bimodality, although we would hope that our indicators are powerful enough so that we can detect a latent taxon by means of *some* features of the distribution, when the multivariate distribution in the phenotypic hyperspace is analyzed by appropriate taxometric methods.

The strongest meaning of "taxonicity" in causal terms is that the disease entity or personality type has its causal origin in a *specific dichotomous etiological agent* such as a mutated gene, or a specific germ, or a particular life history event, for example, a head injury In the ideal case, the specific causal agent C is nonarbitrarily dichotomous, that is, it is present or absent on the basis of a predicate or property that does not possess intermediate gray region degrees. (Perhaps it "could," conceptually, but in fact it does not.) The strongest meaning is exemplified when the dichotomous etiological agent C is both necessary and sufficient for the phenotypic disease syndrome to appear, as in Huntington's Disease, in which all those who carry the Huntington mutation will develop the symptoms provided they survive the morbidity risk period, although in some Huntington sibships this means that a person might have to live to be 70 years of age. A slightly weaker meaning, but one that we find readily accepted by everyone as inherently taxonic, is that of a sine qua non, a necessary but not sufficient condition for the syndrome (Meehl, 1972c). In medicine and genetics, *threshold effects* and *step functions* are also strong meanings of taxonicity. Thus one who receives less than a specified minimum intake of niacin for a sufficient time period will develop pellagra.

Once we pass beyond these forms—necessary and sufficient dichotomous etiology, necessary dichotomous etiology, and necessary threshold (or step-function) etiology—we are in a region of decreasing strengths of specific influence, in which it would be easy to find disagreement among reflective and informed persons as to whether they would consider them instances of specific etiology or not. For these other forms of strong influence mathematically defined, see Meehl (1977). The easiest way in this frame of reference to avoid semantic disputes about whether a situation should be called truly taxonic or not is simply to specify which of the kinds of strong influence the theoretician has in mind.

It is a puzzling historical fact (Meehl, 1979) that formal cluster algorithms, that is, taxometrics of the kind we have earlier called Type II, the classical cluster analytic problem, have not been responsible for discovering a single taxon in psychopathology or, so far as we are aware, in organic medicine. In fact, one cannot even make a clear case for saying that while informal clinical (or experimental) methods have initially revealed the entities, formal cluster algorithms have at least clarified subsequent taxonomic controversies. Should an entity be subdivided? Ought certain atypical clinical cases be subsumed under an entity as *formes fruste*? Is an entity genetically homogeneous? Should a taxon be extended; for example, is Hoch and Polatin's "pseudo-neurotic schizophrenia" concept useful (Hoch & Polalin, 1949), or is it an illegitimate and confusing extension of

the concept schizophrenia?

Alas, even such a clarifying function as to entities provisionally conjectured from nonformal methods by clinical experience cannot be historically documented. Meehl (1979) has in the above cited article listed eight possible explanations of this somewhat surprising state of affairs, which, if left unexplained on the shelf, might reasonably discourage the psychoclinician from pursuing taxometrics, since it doesn't seem to have much payoff even if it's philosophically and mathematically amusing. We have no stake here in denigrating formal cluster algorithms, but content ourselves with that observation, referring the reader to Meehl's paper cited and moving now directly to that one of the three taxometric knowledge situations to which we have devoted our efforts, namely, Type III, the conjectured latent taxon problem.

THE BOOTSTRAPS APPROACH

The first knowledge situation, that is, the case in which we have an accepted criterion and the investigator, knowing there is a taxon, wishes to do a better job identifying the persons who belong to it, is outside the scope of this chapter.

For Case I, where there is an accepted criterion variable telling us to which taxonomic class each individual actually belongs, there are well-known methods, such as linear discriminant function analysis first developed by R. A. Fisher (1936). For [this *Handbook*] to be complete, it should include a chapter describing these methods, and we originally prepared this chapter with that understanding (see Preface [to *Handbook*]). Here we refer the reader to only some of the textbooks and papers that the quantitative researcher in psychology will find useful. In order to be consistent with the usual terminology in this literature, we refer to the indicator variables (e.g., signs and symptoms) as the "independent" variables and the virtually infallible criterion variable as the "dependent" variable. The general mathematical-statistics problem is to find the function of the independent variables that gives the most accurate classification on (or, more generally, maximizes the statistical association with) the dependent variable. Probably the most commonly used solution to this problem is for *linear* functions of independent variables such as in the discriminant function and multiple regression methods. Several appropriate textbooks discuss different versions of these and related methods (e.g., Lindgren, 1962; Draper & Smith, 1966; Rozeboom, 1966; Cornfield, 1967; Tatsuoka, 1971; Overall & Klett, 1972; Kerlinger & Pedhagur, 1973; Finn, 1974; Lachenbruch, 1975; Morrison, 1976). The desirability of using differential rather than equal unit weights in the linear function should not be taken for granted, since the former are subject to sampling error so that the linear function sometimes lacks robustness when used in samples other than the one on which it was developed (e.g., see Guion, 1965; Nunnally, 1967; Darlington, 1968; Wainer, 1976; Dawes, 1979; Dawes & Corrigan, 1974). When using a weighted linear function, there are useful papers regarding missing data (Chan & Dunn, 1972), estimation of classification error rates in new samples (Lachenbruch & Mickey, 1967; Dunn & Varady, 1966; Dunn, 1971), and unreliability in the dependent variables (Gilbert, 1969; Goldstein, 1977).

Regarding the use of dichotomous independent variables, Lachenbruch and

Mickey (1968) show that the "jackknife" method, where each individual is classified by a different discriminant function (that derived on the remaining $N-1$ individuals in the sample) is superior to several other methods in that it yields (almost) unbiased estimates of error rates.

The advantages of using a function more complex than linear to capture interaction, configural, and nonadditive effects were suggested long ago (e.g., see Meehl, 1950) and studied (e.g., see Goldberg, 1969). The use of a function-free actuarial table was advocated by Lykken (1956) and Lykken and Rose (1963). When the independent variables are dichotomous (or categorical), Bayes' Theorem (e.g., see Hays, 1973) and applications of it (e.g., see Solomon, 1961; Bailey, 1965) are the most obvious nonlinear method.

The above references are in no sense complete and in part represent the result of our own selective sampling of the vast literature on Case I research methods.

We will briefly survey the formal cluster algorithm methods used for the classical cluster analytic problem and then devote the remainder of the mathematical discussion to Case III, the conjectured latent taxon problem, after saying something about bootstraps methods generally.

The term "bootstraps effect" was first introduced by Cronbach and Meehl (1955). Surprisingly, the terminology was criticized by some intelligent people on the grounds that it implied that you were lifting yourself by your bootstraps; whereas the reason Cronbach and Meehl chose that terminology was precisely to emphasize the counterintuitive finding that one can psychometrically "lift oneself by the bootstraps," if it is done right. The essential idea of the bootstraps effect as described by Cronbach and Meehl in 1955 was that one can construct an empirical scale such as an MMPI key by empirical keying against a rather poor, low-accuracy fallible criterion (such as unreliable diagnoses by unskilled psychiatrists) and, if one is lucky and clever and the item pool is rich enough, can nevertheless emerge with a scale that diagnoses more accurately than the clinical examiners did. It is incredible but true that there are still psychologists who don't understand this simple point. Bootstraps effects abound in the other sciences, and for that matter, in ordinary human artifacts. Thus, for instance, humankind began making tools with chipped flint, and it is paradoxical but true that we now can grind metal surfaces to a smoothness that involves a variation of only a few molecules' thickness.

A simple example of a bootstraps effect is a disease entity like general paresis. As late as the turn of the century, it was still argued whether or not lues was the specific etiology of "general paralysis of the insane," and the argument was not clinched until 1913 when Noguchi and Moore found the spirochete in the brains of paretics. But notice that in order to find the spirochete in the brains of paretics, one had to rely upon a fallible, although by then rather high accuracy, diagnosis of who is a paretic. Of course there was an intermediate finding of the characteristic cerebral *pathology* (without the etiology), so Noguchi and Moore did not have to rely solely upon phenotypic phenomena like the Argyll-Robertson pupil, or the patient spilling soup on his vest, showing poor judgment and irritability, having trouble saying "hard riding artillery brigade," and so forth. But suppose the biochemist or clinical neurologist had done as we sometimes are forced to do in

psychometrics, that is, had tried a large number of *candidate* indicators on paretics versus patients suffering from various other kinds of neurological disorders also involving impairment of motor function and a dementing process. Suppose that the diagnosis of paresis at 1875 was only 75% accurate in the eyes of Omniscient Jones. If the neurologist had tried out a long list of tests, determining blood turpentine level and the Wasserman and the spinal fluid gold colloid test and the Schick test and the Mantoux, what would happen? Positive blood Wassermann and first zone gold colloid spinal fluid would each have an extremely high construct validity for paresis. An "item analysis" of a long list of candidate indicators would select these two items (both in reality 95% or better two-way pathognomic for the disease entity), even though they had to be discovered on diagnoses that were only 75% accurate, and hence cannot *display* their true validity against formal clinical diagnosis as the available criterion.

Of course the problem is to know when you have successfully achieved a bootstraps effect of this sort; but it's obvious that medicine has done it repeatedly, as have other scientific disciplines. So there is nothing all that strange about the bootstraps effect in psychometrics, except that psychologists are so hooked on operationism that they have been reluctant to admit it. The best example of it in psychometrics is doubtless the intelligence test, which was originally seen to be valid, in contrast with previous efforts to measure general intelligence with Wundtian procedures like two-point touch threshold, because it showed a characteristic developmental change with age, and it agreed well with the pooled judgments of schoolteachers as reflected in social facts like age-grade location.

Consider another example outside psychometrics. If three surveyors each come forward with new methods for estimating how far away a distant mountaintop is, and each of them offers a plausible, but not cogent, theoretical rationale for thinking his method works, then a discovery that the three methods converge in their estimates to a high numerical accuracy tends to corroborate the construct validity of all three of them. For this reason the common undergraduate principle, "Unreliability proves lack of validity, but reliability doesn't prove validity," is somewhat misleading because a nonchance agreement of anything, including a pair of test items or two forms of a test, does prove validity *for something or other*. If that kind of "convergent estimate" reasoning were not essentially valid, no science, including the physical sciences, could get off the ground.

The most dramatic example of bootstrapsing we know of in taxometrics is the "Super-Bootstraps Theorem" proved by Meehl in 1965 (pp. 37-48; cf. also Meehl 1973, pp. 216-217). Suppose that one has a set of fallible phenotypic indicators of a latent taxon that are substantially independent within the taxon and its complement, this being checkable by several nonredundant consistency tests (see below). Having estimated the taxon base rate P in a given population, we define subgroups of patients on the basis of the pattern of positive and negative indicators. Thus, if there are three fallible indicators, each cut at a certain point to maximize hits, we have $2^3 = 8$ cells of patients, within each of which a Bayes' Theorem inverse probability of taxon membership is computable from the patient's indicator pattern. The Super-Bootstraps Theorem shows that if the latent structural model is satisfied and the sample is large enough to give stable estimates

of the numerical values, then, if there exists an optimal cut on some new fallible indicator *v*, chosen so that the proportion of patients classified as within the taxon of interest matches the taxon membership probability of any one of the cells, in order for this cut to yield v^+ rates that match the Bayes' Theorem probabilities in the remaining cells, it is *both necessary and sufficient that the v-cut be infallible* (two-way pathognomonic, functioning perfectly both as an inclusion and exclusion test). This is such a remarkable theorem that we have been puzzled why, despite a generalization of it published by Dawes and Meehl (1966), it has come in for almost no attention in taxometrics.

MAXCOV-HITMAX METHOD

We illustrate our conception of bootstraps for taxometric Case III, the conjectured latent taxon problem, with a simple crude form of one of our taxometric methods, MAXCOV-HITMAX (Meehl, 1973a, pp. 200-224). This is only for pedagogic illustrative purposes here, as a more general and rigorous formulation of the MAXCOV-HITMAX Method will be explained later on. Suppose we conjecture that in a population of psychiatric patients not formally diagnosed as schizophrenic, there exists a taxon, the *schiozotype* (Rado, 1956; 1960; Rado & Daniels, 1956; Meehl, 1962; 1972a, 1972c). Our theory sketch, based upon some combination of clinical experience, psychiatric genetics, and preliminary statistical analysis of file data, is that there are quite a few patients who have the schizotypal personality organization (and, on our dominant gene hypothesis of the disorder, are carriers of the specific schizogene) but who have not decompensated clinically to the extent of presenting the classical textbook picture of a "florid schizophrenia." These considerations from our antecedent state of knowledge are mentioned solely to motivate the research strategy. They are not going to be *relied on* in the traditional statistician's sense of "needful assumptions." We do not treat any of the conjectures, main or auxiliary, as "assumptions" in that strong sense. Everything said, whether in the formalism or the interpretative text, is conjectural, although obviously at any given stage of scientific and clinical knowledge of a disease entity or personality type, some conjectures are more problematic than others. That a conjecture is highly problematic, in the sense of taking a big risk, going out on a limb far beyond the facts presently known, does not speak against it. From a neo-Popperian standpoint this is, on the contrary, a methodological plus. We are interested in trying to test this risky conjecture that a taxon "schizotype" really exists in our clinic population.

Note how much is conjectural and how little corroborated at this point. We conjecture that there is such a taxon, but we do not "know it." Philosophically, of course, one never knows anything empirical for certain, but in common sense and in most scientific usage, highly corroborated conjectures that also fit into a larger network of highly corroborated theoretical conceptions are said to be "known." If there is such a taxon, we do not know with what relative frequency it occurs in our population, that is, we are ignorant of its base rate *P*. We do not know whether the extra-taxon class (*taxon complement*) is itself taxonically differentiated, although we think it likely, on general previous knowledge, that it is; but taxometric discrimination among the cases outside the conjectured schizotypal taxon is not of

interest to us except as a dangerous source of error in identifying the schizotype. We have tentative notions about some indicators (interview, psychometric, neurological) of the taxon, but we lack numerical estimates of their validities and the optimal cut on those that are dimensional rather than a qualitative predicate. We do not know the valid and false positive rates achievable by the best cut. We do not know what the indicator distributions are shaped like, and we are unwilling to conjecture that they are normal or equal in variance. We do not know whether the candidate indicators are appreciably correlated within the schizotypal taxon, if it exists, or outside it. We are aware of the unpleasant possibility that a schizotypal discriminator may also have some validity for some unidentified extra-taxon groups, in which case the indicators will necessarily be correlated pairwise in the complement. In short, *we know essentially nothing*, and about some of these quantitative questions *we do not even have plausible conjectures.* Such a knowledge situation is evidently a classic case appropriate for bootstraps taxometrics. How can we get taxometric knowledge out of ignorance? We shall see.

Suppose we have three quantitative indicator variables conjectured to have some validity for the schizotypal taxon, namely, x: MMPI Scale 8 [$= S_c$]; y: a Rorschach schizoid composite index; and z: quantified subclinical dysdiadochokinesia. The choice of indicators is motivated partly by previous experience suggesting validity, and partly by the desire to choose phenotypically "nonoverlapping" behavior domains to minimize methods covariance, so that the conjectures of near-zero correlation within the taxon and within the complement have a good chance of being nearly correct. In addition, we may undertake a minimizing of undesired intrataxon covariance by ordinary item analytic procedures (done on diagnosed groups) for those variables that are item composites, such as structured inventory scores, or in the case of nonitem measures, by a suitable nonlinear transformation of the metric.

If desired, a preliminary investigation of this intrataxon correlation question can be conducted, although the interlocking consistency tests discussed below should suffice to corroborate that auxiliary conjecture. The problem about direct testing of the near-zero correlation auxiliary conjecture is, obviously, that in a mixed psychiatric population, there is some unknown base rate P of schizotypy. *Therefore, cases categorized by formal diagnosis cannot be used to answer this question by direct empirical methods.* However, we can increase the prior probability of that auxiliary conjecture by testing not the mixed bag of psychiatric patients but a group of thoroughly studied patients diagnosed as clinically schizophrenic. Ideally, we would add some cases diagnosed "schizophrenia in remission" and, in some prospective studies, preschizophrenics (e.g., college students who took the MMPI as freshmen and subsequently are found in state hospitals with a diagnosis of schizophrenia). If the pairwise correlations of the three indicators x, y, z are close to zero in various diagnosed schizophrenic populations, we have some confidence that a semicompensated (e.g., pseudo-neurotic, borderline state) psychiatric population would also show low correlations, although that does not follow with necessity.

On the other side, consider a sample of nonpsychiatric presumed "normals." On our conjecture that the specific etiology is a dominant schizogene of low

penetrance, the fact that the diagnosed schizophrenia rate in first-degree relatives of schizophrenic probands is in the neighborhood of 12%, instead of the theoretical 50% on a perfectly penetrant dominant gene hypothesis, means that the clinical penetrance is only about ¼. Since the general population lifetime risk for schizophrenia is in the neighborhood of 1%, this suggests that the frequency of persons in the general nonpsychiatric population carrying the schizogene is around 4%. This being a rather small proportion, we can safely assume that the statistical tail can't wag the normal population dog enough to generate a taxonically induced covariance of *x, y,* and *z* among normal persons via the presence of only $p = .04$ of schizotypes among them. Therefore, the auxiliary conjecture in the complement class of zero correlation is fairly directly testable by correlating *x, y,* and *z* pairwise in a normal population. We repeat, these preliminary investigations are worth pursuing from the standpoint of saving time by early discovery of indicators that are excessively correlated within the taxon or the complement, but they are not strictly necessary because the auxiliaries will be part of the network tested by the consistency tests.

We begin with a general formula for the covariance of a mixed population, where it is an algebraic truth (not dependent upon distribution assumptions) that the covariance of a mixed population is a simple sum of three terms. The schizotypes are represented by subscript *t* (= taxon), the nonschizotypes by subscript *c* (= complement), and the base rate of schizotypy *p.* Then the (*yz*) covariance for a mixed population is given by

$$\mathrm{cov}(yz) = p\,\mathrm{cov}_t(yz) + q\,\mathrm{cov}_c(yz) + pq(\overline{y}_t - \overline{y}_c)(\overline{z}_t - \overline{z}_c) \qquad (1)$$

This expression is derived in Meehl (1965, pp. 12, 28-29; see also Meehl, 1968, pp. 4-5). It is intuitively obvious that the covariance of a mixed group should depend partly upon the two within group covariances, these components being weighted in proportion to the group rates; and the third term, a kind of "validity-mixture" term, reflects the influence of the crude validities expressed as differences of the latent means between the schizotypal and nonschizotypal subpopulation on the two indicators, weighted by the product term *p q,* which is a measure of taxonic "mix." One way to look at this third term intuitively is that patients high on indicator *y* will tend to be schizotypes, and therefore, if indicator *z* has validity, they will tend to be high on *z*. That is, *the covariance is generated by the fact of taxon mixture.* So this third term is the interesting one for taxometric purposes. We see that it tends to increase with increases in the true indicator validities, that is, the differences $(\overline{y}_t - \overline{y}_c)$ and $(\overline{z}_t - \overline{z}_c)$, and also with the degree of taxonic "mix," measured by the cross product of the schizoid taxon's rate *p* and its complement $q = 1 - p$. This accords with intuition, since if the indicators had zero validity for the taxon, or if they had taxonic validity but the taxon was unrepresented in a population subjected to statistical study, this third term would vanish.

The greatest taxonic "mixture," that is, the opposite extreme from a "pure" population (consisting only of the schizoid taxon, or solely of individuals not belonging to it, where the cross product term $pq = 0$) is that of an even mix. For fixed crude indicator validities $\Delta\overline{y} = (\overline{y}_t - \overline{y}_c)$ and $\Delta\overline{z} = (\overline{z}_t - \overline{z}_c)$, the smallest values of the "validity-mixture" term of the general expression are those for $p = 0$, $q = 1$,

or p = 1, $q = 0$; and the greatest value of this interesting term occurs for the even mix $p = q = ½$, where the product is equal to ¼. That is, in a population in which half of the patients were members of the schizoid taxon and half were not, the expected value of the validity-mixture term would be ¼ the product of the latent mean differences, that is $¼\,\Delta\bar{y}\Delta\bar{z}$.

Now making the simplifying approximate assumption (better, auxiliary conjecture) that the indicators have been chosen or constructed in one of the three ways indicated above, so that within the schizoid taxon each indicator correlates negligibly with each other one, and ditto for the complement class, then in our equation

$$\mathrm{cov}(yz) = p\,\mathrm{cov}_t(yz) + (1-p)\,\mathrm{cov}_c(yz) + pq\Delta\bar{y}\Delta\bar{z} \tag{2}$$

the first two terms vanish. Hence the total observed covariance cov(yz) is simply the product of the schizoid and nonschizoid rates times the product of the crude latent validities. The observed covariance of such a mixed population is contributed solely by the indicator validities and the fact of taxonic mix.

In Figure 5.1 we have a diagram with the vertical axis being a measure of clerical ability, and the horizontal being a measure of *n Nurturance.* Assume that, for the female sex considered by itself, there is no correlation between clerical ability and *n Nurturance.* Then the bivariate frequency distribution for females can be drawn as a circle, and the regression line of y on z within that circle is flat, has slope = 0. Overlapping with that circle representing the females, but southwest of it

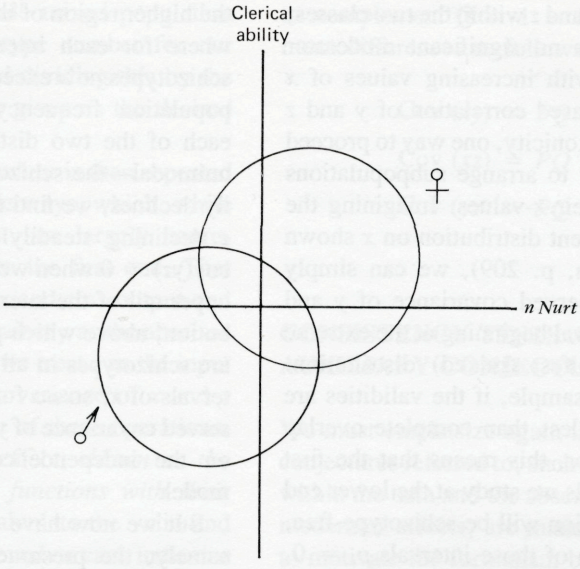

Figure 5.1. Illustration of two indicator variables independent within two taxa.

on the graph, we see another circle representing the (yz)-distribution for males, assuming that within the male sex also there is no correlation between clerical ability and *n Nurturance.* This situation is the idealized one for taxometric analysis, because the two indicator variables are completely independent within the two taxa. Now if we consider the mixed population of males and females,

these two overlapping circles, the regression line within each of which is flat, are fused as an ellipse elongated in the northeast direction. The best fitting regression line for the mixed population does not have 0 slope but a positive slope, depending upon the size of the sex difference in clerical ability and *n Nurturance*

If there were some way of arranging a sequence of clinical subpopulations in the order of their schizotypal rates, then, on the auxiliary conjecture that the covariances are zero within the taxon and its complement, the observed covariances cov(*yz*) of that ordered sequence of populations should commence at zero for the subpopulation that contains no representatives of the schizotypal taxon and should steadily increase to a maximum, achieved within the subpopulation with $p = q = ½$, the maximum taxonic mix in which half the patients are schizotypal and half are not; and then should decline steadily to reach and remain at zero when we move into subpopulations in which all of the patients are schizotypal. This relationship between the observed covariance within a population of unknown but conjectured mix is already interesting and could be used for sign validation purposes, and to test the model, if we had before us any group of populations somehow identifiable as differing in the schizoid rate (see, e.g., Dawes & Meehl, 1966).

But we need not know how to identify such populations qualitatively, nor the successive increments in schizotypal base rate in moving from one to the other, in order to carry out a remarkably powerful bootstraps operation, as follows: We have conjectured that a third indicator *x,* not yet used, also has taxometric validity. Assuming that *x* is uncorrelated with *y* and *z* within the two classes, and that there is no significant moderator effect, so that with increasing values of *x* there is an increased correlation of *y* and *z* apart from the taxonicity, one way to proceed would be simply to arrange subpopulations as defined by their *x*-values. Imagining the picture of the latent distribution on *x* shown in Meehl (1973a, p. 209), we can simply compute the observed covariance of *y* and *z* for each *x* interval beginning at the extreme left of the manifest (mixed) distribution. With a sizeable sample, if the validities are such as to have less than complete overlap on the *x*-indicator, this means that the first few class intervals we study at the lower end of the *x* distribution will be schizotype-free. That is, for each of those intervals $p_i = 0$, and therefore the covariance of the two "output" indicators $\text{cov}_i(yz)$ will be at zero within all of those intervals. Then as we begin to pick up a few schizotypes in the low tail of the latent schizotypal frequency function so that $p > 0$ and increases steadily while *q* goes down from 1, the observed covariance of *yz* takes on nonzero values and begins to increase. The hitmax interval on *x* is the interval above the intersection of the latent frequency functions on *x.* Cutting there maximizes the hit rate, taking account of the base rate *P,* so that the functions we are dealing with here should be unrelativized frequency functions, their ordinates being not intrataxon density functions but frequency functions reflecting the base rates *P, Q.*

The hitmax interval h_i we do not observe directly, because we do not know the latent distributions including their base rates, and so on, and we have no accepted criterion. But h_i is nevertheless locatable by this bootstraps procedure. It will correspond to the class interval on the input variable *x,* which yields the maximum value of the output variable covariance $\text{cov}_i(yz)$, that being the interval for which

$p = q = ½$, that is, for which the taxon mixture term has its maximum ($= ¼ \Delta\bar{y}\Delta\bar{z}$). As we pass through the hitmax interval on indicator x and move into the higher region of the mixture distribution, where for each interval the proportion of schizotypes now exceeds one-half of the subpopulation frequency, and for which—if each of the two distributions is separately unimodal—the schizotype proportion steadily declines, we find the covariance of y and z declining steadily. It reaches the value $\text{cov}(yz) = 0$ when we come to the end of the upper tail of the latent nonschizotype distribution, above which point all of the patients are schizotypes in all of the higher class intervals of x, so no further change in the observed covariance of y and z is to be expected on the independence assumptions of the model

But we now have a very useful number, namely, the presumed constant product $K = (\Delta\bar{y}\Delta\bar{z})$ of the crude latent validity increments. We know that for cases lying in the hitmax interval, $\text{cov}(yz) = ¼ \Delta\bar{y}\Delta\bar{z} = ¼K$, the constant K being the fixed product of the crude "validities" (latent mean differences), permitting us to solve for K. Further, the general expression for the covariance of a mixed group holds for each of the class intervals, not just for the hitmax interval. Now that we are able to estimate the latent constant K by means of the hitmax interval's observed (yz)-covariance, we can plug in that parameter K in the general expression $\text{cov}_i(yz) = p_i q_i K$ for each x-interval. So in each x-interval we have a quadratic in p,

$$Kp^2 - Kp + \text{cov}_i(yz) = 0 \tag{3}$$

which we solve to obtain p_i in each x_i-interval. Of the two roots, which will of course be fractions, one less and one greater than ½, we choose the smaller root for all the x-intervals below the hitmax cut, and the larger root for all the x-intervals above the hitmax cut.

Now we are in excellent shape, since by multiplying the observed crude frequency in each x_i-interval by the schizotypal rate p_i of that interval, we obtain the number N_i of schizotypes in each class interval of x. Adding these over the whole mixed population yields the total schizotype number for our sample $\Sigma N_i = N_s$, and dividing this by the grand N of our sample gives us the desired latent base rate P.

This bootstraps procedure is then repeated in the two other possible ways, which, while not totally independent, are largely nonredundant with respect to the first one. That is, we treat indicators x and z as output and plot their covariance against y as the abscissa variable, and similarly we treat z as the input variable and plot the covariance of x and y as output. In the course of these operations, the computer has, in effect, *drawn us the three latent frequency functions with their overlaps*. So we can calculate the valid and false positive rates achieved on each indicator taken separately when patients falling above the hitmax cut of that indicator are classified as schizotypal and those falling below the hitmax cut as nonschizotypal.

Finally, knowing the grand base rate P and the valid and false positive rates, p_{tx}, p_{ty}, p_{tz}, p_{cx}, p_{cy}, p_{cz}, we can plug these numerical values directly into Bayes' Formula for inverse probability. Thus, suppose we have a patient from this

population who falls above the hitmax cut on indicators x and y, but below the hitmax cut on z. This pattern (x^+, y^+, z^-) of scores on the three indicators yields a Bayes' Theorem inverse probability that the patient is schizotypal as follows:

$$P(Sc / x^+ y^+ z^-) = \frac{P p_{tx} p_{ty} q_{tz}}{P p_{tx} p_{ty} q_{tz} + Q p_{cx} p_{cy} q_{cz}} \tag{4}$$

We present here only one example of a consistency test, namely, the total covariance test. Since the general equation for observed covariance of a mixed population holds generally, it holds for the total population under study, as well as for any of the subpopulations, however identified. Consequently, the grand covariances of x and z, of y and z, and of x and y for the whole (mixed) population can be written in terms of the base rate P, its complement $Q = 1 - P$, and the latent mean differences as follows:

$$\begin{aligned} \mathrm{Cov}(xy) &= PQ\,\Delta\bar{x}\Delta\bar{y} \\ \mathrm{Cov}(xz) &= PQ\,\Delta\bar{x}\Delta\bar{z} \\ \mathrm{Cov}(yz) &= PQ\,\Delta\bar{y}\Delta\bar{z} \end{aligned} \tag{5}$$

CONSISTENCY TESTS AND AUXILIARY CONJECTURES

We must emphasize again that the auxiliary conjectures referred to, such as independence within the taxa and the absence of significant moderator effects, are made as assumptions to motivate the formalism but are not being empirically *relied upon* as trustworthy (or pious hope!) postulates. Rather they are jointly tested, together with all of the other auxiliary conjectures and the main assumptions of the model, by imposing numerical tolerance limits on departures from the theorems derived as consistency tests. The reader will totally misunderstand our bootstraps approach, and the neo-Popperian philosophy of science underlying it, and will consequently not know how seriously he or she should take the formalism, if he or she does not grasp our approach to this question. The satisfaction (within tolerances) of the several consistency tests, which are not equally responsive to departures from different aspects of the conjectural latent model, shows what? It corroborates the substantive theory that there is a taxon, and that these indicators have substantial validity for it, together with all of the other conjectural assumptions in the explanatory text that interpret the formalism or justify taking a given step within it. This means that we are testing the crude theory sketch of the schizoid taxon *and* the auxiliary conjectures; *and* we are also tightening up the network defining the loose concept, "schizotaxia," psychometrically; *and* we are also validating the indicators. All of these are done simultaneously. When we corroborate the consequences of the nomological network, we corroborate the network.

We do not dispute the desirability of being able to test substantive conjectures under circumstances in which the auxiliary conjectures are unproblematic. The difficulty is that in the life sciences, and certainly in the behavior sciences, this is an unusual state of knowledge for the investigator to be in when pursuing taxometric research. Example: In testing a dominant gene theory of schizotypy by

studying the Rorschach responses of parent pairs of diagnosed decompensated schizotypal probands ("clinical schizophrenia"), we may rely upon psychometric extrapolation of Rorschach indicators from validation studies. But these studies are based not upon the compensated relatives of known schizophrenics but upon patients who themselves have been schizophrenic, or, very rarely, who were tested when normal and subsequently became schizophrenic. While this extrapolation has a certain plausibility, it is obviously not certain. An investigator who had immersed himself or herself in the genetics of schizophrenia might well surmise, at a given stage of knowledge, that the dominant schizogene theory is itself no more problematic than the Rorschach psychometric sign he or she plans to use in testing that theory. For a further discussion of this, see Meehl (1978) and a forthcoming book by Golden and Meehl. We must emphasize that *qualitatively* the situation in the life sciences is not different from that in the inorganic disciplines like astronomy and chemistry, in which breakthroughs have sometimes been achieved by challenging auxiliary assumptions that had been considered well corroborated. But there is a big *quantitative* difference, and it is foolish for the behavior taxometrist to pretend otherwise.

We do not wish to defend an orthodox Popperian position as to the inherent desirability of low probability substantive conjectures (Grünbaum, 1976). But we think it fairly clear from the history of the sciences that theories do get their best support in the eyes of the scientific community by virtue of having made risky predictions. Our own view is that scientists usually prefer the theory *T* itself to be antecedently probable on the background knowledge, including other "accepted" theories. But they want its observational consequences to be antecedently improbable, *absent the theory.* This combination goes through nicely on a Bayes' Theorem representation, but the logician then has to worry about how *T* can be antecedently probable and its observational consequence-class not. Be that as it may, as Salmon puts it (1966, p. 119):

> A hypothesis risks falsification by yielding a prediction that is very improbable unless that hypothesis is true. It makes a daring prediction, for it is not likely to come out right unless we have hit upon the correct hypothesis. Confirming instances are not likely to be forthcoming by sheer chance.

Whether on a Bayesian or Popperian or commonsense approach, the scientist gains respect for a theory when it makes a numerical point prediction or any other kind of detailed specification of facts not foreseeable from some alternative theoretical construction. Put simply and without reference to technical philosophy of science controversies still current today, the general idea is that we are impressed when a theory makes detailed predictions that other competing theories don't, *because we think that for a theory to have no truth to it and yet make detailed predictions about facts would be a strange coincidence*

Some such reasoning must underlie all empirical science, however the philosophers reconstruct it, because otherwise everyone could decide whimsically, apart from the strength or weakness of the empirical predictions theories make. Point predictions in which a narrow range of values of an observed variable is specified ("narrow" means with reference to background knowledge as to the order of magnitude of that observational measure's range of variability, theory aside) are

the paradigm case. Other strong corroborators are predictions of rank orders, second order differences, function forms (when more complicated than a straight line and when the points to be fitted are considerably in excess of the number of parameters adjusted), and the like (see Meehl, 1978, pp. 824-825)

One of the strongest kinds of support for any scientific concept is an inference to a numerical point value on the basis of a theoretical model in which the numerical value is independently arrived at through manipulations of the observational facts within a formalism that would have no rationale were it not for the conjectured latent causal or structural model of the theory. If we get the same answers when each numerical prediction took a risk, this provides very strong corroboration. Thus, for instance, physical chemists believe that we know the Avogadro constant *N,* the number of molecules in a gram molecular weight of a substance, with exactitude, because there are some 14 non-redundant ways of estimating it, ranging from the properties of soap films to purely theoretical considerations based upon quantum mechanics. Convergence of numerical point estimates is therefore to be sought wherever possible, and taxometric bootstraps operations are a good place to do this.

We have not found it possible to make a clear formal distinction between main estimators and consistency tests. A given expression into which numerical values can be plugged, or from which numerical values can be estimated when we have assigned all but one of the values in the expression, can evidently serve either as a main estimator of a latent parameter or as a consistency test. If we have more equations than unknowns, we have consistency tests. If we have fewer equations than unknowns, we have neither main estimators or a consistency test. If the number of equations equals the number of unknowns, then we have estimators but the system cannot be inconsistent, because empirically that will never happen. These matters are better discussed in the context of presenting the more detailed and generalized formalism below

APPROXIMATIONS AND VERISIMILITUDE

It is important that the reader understand why we are not much interested in developing exact statistical significance tests for consistency formulas. We could say gently that it has low priority among the current research tasks of a taxometrist. But more honestly, we view it as largely a waste of time, springing from a misleading and counterproductive philosophy of science. An exact studentized significance test for a consistency test of the taxometric model would serve the purpose of refuting the theoretical latent model as taken literally. That is, if the consistency test is "failed," meaning that its numerical values have a low probability α of occurring on the theoretical hypothesis stated in the formalism, what this shows is that the functional relations stated in the formalism, such as complete independence, absence of moderator effects, continuity, and so on, are not literally true of the state of nature. *We do not understand why anyone would want to show that, since we already know that in advance.* Whether or not this α level is achieved will depend jointly, and in a complicated way (which nobody knows how to set out rigorously and analytically in the mathematics), upon the degree of departure of the state of nature from the idealized model as conceptualized in the

embedding text that exposits and interprets the formalism, together with the inherent idealizations represented in any formalism, as for example, taking derivatives. In standard analysis differentiation involves assumptions of continuity about variables that must of course be discontinuous in the life sciences—certainly when we add finite numbers of test items in a psychometric score! The most important factor involved here other than the degree of distortion of the truth by the model is the power function of the significance test. Nobody could safely contradict the following assertion about taxometrics in the behavior sciences: Given a sufficiently large sample, all conceptualizations of latent taxa that cause certain indicators to covary in specified ways as stated in the language of differential calculus is literally false. Consequently whether or not an investigator succeeds in *showing* it to be literally false depends solely upon the power function, and therefore mainly on sample size (see Meehl, 1967, 1978, and more generally, Morrison & Henkel, 1970).

Of course this does not permit the taxometrist to be intellectually irresponsible. On the contrary, reflection on the uselessness of significance tests in these contexts leads to the desirability of constructing models of approximation theory, an important matter that has been insufficiently pursued in the social sciences. The problem is never one of literal truth of the conjectural model as stated in the formalism and the embedding text, since those are always false, taken literally. The problem is the setting up of tolerances in the light of robustness considerations, and as yet only limited inroads have been made into that problem.

Finally, we do not mistakenly suppose that the neo-Popperian concept "verisimilitude" is in satisfactory shape in contemporary philosophy of science. But we are convinced that *some such* metatheoretical concept is unavoidable, since all theories are false, being idealizations. A theory of schizotypy could have high verisimilitude, although it erred in regard to the absence of nuisance variables operating within the taxon to generate a certain correlation intrataxonically, violating the independence assumption in our derivation above. But that would be a small departure from the truth compared to the situation in which there is simply no such thing as the schizoid taxon at all, or one in which the indicators, while valid for the genetic schizoid taxon, also have high validity for a completely environmentally determined phenocopy. Science cannot wait in the development of its instruments upon a completely satisfactory philosophical analysis of verisimilitude, any more than historically it has had to wait upon an adequate clarification of the notions of induction, proof, or even for that matter the core concept, truth, itself.

A TAXOMETRIC PROBLEM: TESTING A THEORY OF SCHIZOPHRENIA

Our work in taxometrics began with a desire to test a substantive theory of schizophrenia developed by Meehl. This theory conjectures that the genetic predisposition for schizophrenia is taxonic in nature. After briefly presenting the major features of this theory (for a complete description see Meehl, 1962 [and Meehl, 1990]), we will describe each of three taxometric methods we have developed for testing this theory. These taxometric methods are general in that they can be used

for testing any theory regarding a conjectured taxon. After discussing the methods, we will report the results of applying them in a preliminary' empirical test of the schizophrenia theory.

Twenty years ago Meehl (1962) proposed a theory of schizophrenia in which he hypothesized that only a certain class of people, those with a particular genetic constitution, have *any* liability for schizophrenia. This hypothetical class will be referred to here as the "schizoid taxon." While Meehl's theory has generated considerable interest, there has been little empirical evidence to either corroborate or refute it, probably because the taxonomic nature of the theory resulted in methodological and statistical problems that could not be adequately handled by methods that existed at that time.

A description of the general nature of methodological and statistical problems encountered in testing this theory was given by Meehl (1973a). If, as Meehl proposes, the specific etiology of the schizoid taxon (schizoidia) is a single dominant gene, and the only indicators available are highly fallible phenotypic ones, how can the probability that a person carries this gene be estimated? Currently there is no generally acceptable criterion variable and no definitive diagnostic touchstone, sign, symptom, or trait that can be measured reliably. Not all of the correlates of schizoidia are sufficiently pathological to be called "symptoms" or valid enough to be called "signs" (hence we will use the term "indicator"). Thought disorder or "cognitive slippage," which is viewed by Meehl (following Bleuler) as the primary indicator of schizoidia, is not sufficient by itself for taxonomic purposes. Some clinical manifestations of cognitive slippage can be noted during intensive psychotherapy of psychiatric patients, and can be used as *inclusion* tests for schizoidia, in that their presence is an almost infallible indication of the presence of this particular pathology.

Unfortunately, these manifestations are too rare to be used successfully as *exclusion* tests, their absence does not necessarily imply absence of schizoidia. No valid psychometric test of cognitive slippage is available—an especially serious problem for researchers who would like to study schizoidia as a hypothetical taxon that includes both schizophrenics and schizoids (who may never manifest diagnostically psychotic degrees of cognitive slippage). Therefore, we have a perfect example of a bootstraps problem (Cronbach & Meehl, 1955), in which we must start with a fallible set of indicators of unknown validities and hope to end up with accurate estimates of these validities on the basis of some internal statistical relationships among them.

CLUSTER ANALYSIS

The most popular taxometric methods are currently the cluster analytic ones. As discussed above, these methods are intended to solve taxometric problems where we do not have a conjectured taxon. According to Blashfield (1976) six "agglomerative" cluster methods are the most commonly used. Before using these six methods to attempt to detect the schizoid taxon, we decided to subject each to an empirical trial where the underlying taxonomy is known. We will describe the empirical trials after giving a brief description of these agglomerative cluster methods.

Although cluster methods were first proposed in the social sciences by Zubin (1938) and Tryon (1939), general interest in their use paralleled the development of large computers. Now, according to Blashfield (1976), over 100 different cluster methods are found in Anderberg (1973), Bailey (1974), and Everitt (1974). Even though the cluster methods are most frequently used for generating clusters of related variables, they can also be used for generating clusters of similar individuals, as would be necessary in searching for a schizoid taxon.

The agglomerative methods are used in conjunction with a matrix of similarity-values for each pair of individuals' sets of indicator scores. The measure of similarity between the two individuals' indicator scores is usually some kind of correlation or distance in the indicator-hyperspace. From the similarity-values, clusters can be generated by assigning individuals with similar scores to the same cluster. A cluster method is iterative and generates a hierarchical tree, with each level of the tree representing a different clustering called a partition. If there are N individuals, then the first partition consists of $N-1$ clusters, the next of $N-2$ clusters, and so on until the last partition, which consists of two clusters.

Four of the most popular agglomerative methods are called by Blashfield the "single linkage," "complete linkage," "average linkage," and "minimum variance" methods. Two other cluster methods, mathematically related to the average linkage method, are the "centroid" and "median" methods. It has been shown by Lance and Williams (1967) and Wishart (1969) that all six methods can be described in terms of the same algorithm or iterative procedure.

Each of six clustering algorithms can be easily described to give a sense of how each works and how it differs from the others. For example, in the single linkage method, each member of a cluster is more similar to at least one member of that cluster than it is to any member of any other cluster. In the complete linkage method, each member of a cluster is more similar to the most dissimilar member of the same cluster than it is to the most dissimilar member of any other cluster In the average linkage method, each member of a cluster has a greater average similarity with the other members of the same cluster than it does with the members of any other cluster. In the centroid method the members of a cluster have a greater similarity to the centroid of the cluster than they do to the centroid of any other cluster. The centroid of a cluster is the vector of indicator means calculated across the members of the cluster. The median method is similar to the centroid method except that the median of the cluster members is used in place of the centroid. In the minimum variance method, the clusters are formed so that the sum of the squared differences in the similarity-measures across pairs of individuals of the cluster is minimal.

Detection of the Biological Sexes: An Empirical Trial of a Taxometric Method

We have described a theory of schizoidia, taxonomic in nature, which requires an appropriate taxometric method in order to test it. Because cluster analysis is the best known taxometric method, we considered using it. However, before using a cluster method for our purpose, we need to be confident that it will usually produce "accurate" clusters, or that it will rarely produce totally inaccurate or

"spurious" ones that would be mistaken for "real" clusters. For our purposes a set of clusters is "accurate" or has sufficient verisimilitude (Popper, 1962), if it corresponds closely enough to an actual underlying taxonomy of real empirical classes; if the set of clusters does not have such a degree of verisimilitude, it will be said to be spurious. We can perform a simple empirical trial to obtain an idea of the accuracy of some of the cluster methods, and other taxometric methods we will describe, when used to detect real empirical classes such as those considered when testing a typical taxonomic theory in the social sciences.

In one study (Golden & Meehl, 1981) we evaluated six of the more popular cluster methods as to their accuracy in testing a pseudo-taxonomic theory, one that we knew in advance to be correct. In doing this, we attempted to determine which, if any, of these cluster methods are satisfactory for solving the pseudo-problem of detecting the taxonomic variable of biological sex when sex-discriminant Minnesota Multiphasic Personality Inventory (MMPI) items are used as indicators. Taxometric methods that cannot pass this empirical trial are unlikely to be useful in detecting, as but one example, the schizoid taxon with (other) MMPI items. In this approach to evaluating cluster methods we determine how well each method detects a *known* taxon. General use of this approach is difficult in fields such as personality and psychopathology, since there are few known taxa. Fortunately, at least one physical taxon, biological sex, is virtually perfect on the criterion side. Also it is reassuring to note that a scale of MMPI masculine-feminine interest items produces bimodality for a large mixed sample with equal numbers of males and females. We think it likely that many taxa in the social sciences have too much overlap to generate bimodality when psychometric indicators such as MMPI are used. If bimodality obtains, it is generally because the latent taxonic class distributions on the indicator-scale have very little overlap. For example, if the two taxonic class base rates are equal, and the two taxonic class distributions are normal in shape, then bimodality is only discernible when the two means are more than two within-taxon class sigma-units apart (Murphy, 1964).

The MMPI sex discriminant items are described in some detail here, since these items were also used to test each of our own taxometric models to be described below.

The MMPI item-indicators were chosen by comparing two samples of males and females on each of the 550 MMPI items. These two samples consisted of 430 male and 675 female adult psychiatric patients in the University of Minnesota Hospitals. The items were scored 1 for a "female" response and 0 for a "male" response, with "female" and "male" responses determined both by comparing the response proportions of these same male and female samples and by considering the item content (i.e., face validity). The two methods agree perfectly. It was found that 49 items discriminated between the two samples to the extent that the difference in the proportions that scored a 1 was .10 or more. This difference in proportions will be referred to as the "validity" of the item. Of these 49 items, 18 were found to have validities of .30 or more and will be referred to as the "highly discriminant" items. Examples of the highly discriminant items are given below. The letter in parenthesis indicates that a response of true (T) or a response of false (F) is scored as 1; otherwise the response is scored as 0. "I am not afraid of mice"

(F), "I used to like hopscotch" (T), "I used to keep a diary" (T), "I very much like hunting" (F), "I like collecting flowers or growing houseplants" (T), "I would like to be a nurse" (T). The 12 items found to have validities between .20 and .30 are referred to as the "moderately discriminant" items. Examples of moderately discriminant items are: "I like poetry" (T), "I like to cook" (T), "I would like to be a soldier" (F), "If I were an artist, I would like to draw flowers" (T), "I have no fear of spiders" (F). The 21 items found to have validities between .10 and .20 are referred to as the "weakly discriminant" items. Examples of weakly discriminant items are: "I gossip a little at times (T), "Sometimes when I am not feeling well I am cross" (T), "I would like to be a florist" (T), "At times I feel like swearing" (F), "I am certainly lacking in self-confidence" (T), "I am easily downed in an argument" (T), "I like science" (F). Finally, 26 items (scored arbitrarily with regard to sex direction) were selected at random from the remaining 501 items in the MMPI inventory with validity coefficients between –.10 and .10, and are referred to as "nondiscriminant" items. Examples of very weakly or, as we call them, nondiscriminant items are: "I have several times given up doing a thing because I thought too little of my ability" (T), "I have often met people who were supposed to be experts who were no better than I" (F), "At times I have worn myself out by undertaking too much" (F), "My plans have frequently seemed so full of difficulties that I have had to give them up" (T).

Objection has been made that this task is difficult or impossible, because MMPI items are too many steps removed from the sex-determining genome. One simply cannot expect, it is said, that verbal self-reports of interests, attitudes, feelings, and the like will possess sufficiently high construct validity vis-à-vis the XX genotype to permit a bootstraps taxometric identification of the biological taxon, let alone a highly accurate sorting of individual subjects into those taxa by the use of such fallible bootstrapped indicators.

The answer to this criticism is that we know as an empirical fact that it is possible to do so, as the authors of the MMPI did when they had available an external criterion available for empirical keying. Even the old MMPI M_f scale itself achieves between an 85 and 90% accuracy in identifying biological sex. This suffices to show that the net attenuated construct validity of these kinds of verbal items is *not* too poor for the sex identification task we set to it.

We will also present evidence that each of our taxometric methods can be applied in a bootstraps fashion, without knowledge of the criterion membership of the individuals, to infer the biological sex taxonicity, estimate the taxa base rates accurately, and classify individuals with an accuracy of 85 to 90%. Thus we can bootstrap this taxonomy using MMPI items and get a true validity that compares favorably with that achieved by Hathaway and McKinley employing the objective sex membership dichotomy for criterion keying. For further analysis of the methodological fairness of the empirical trial, see Golden and Meehl (1981).

Our position on the fairness of this empirical trial is that the conjecture that a certain taxometric method will usually enable one to detect a taxon, to find the strong items, and to assign weights to them for classifying individuals into the detected taxon or out of it, is strongly discorroborated when the method fails at its task in a context in which the dichotomy is known to exist and the fallible indica-

tors available to the taxometric method are known to be sufficiently valid so that when put together, even by a crude item analytic and unweighted procedure, they are highly accurate.

The Results of the Empirical Trial of the Cluster Methods

The accuracy with which each of six cluster methods detected the biological sex taxonomy was determined by observing how accurately a mixed-sex sample of individuals was classified according to biological sex. The last partition of two clusters was used; that with the most females was identified as the female cluster and the other as the male cluster.

For each trial, a mixed sample of size 200 consisting of 100 males and 100 females was analyzed. Two hundred is a common size for cluster analysis studies and was the maximum that the available computer program could accommodate. A second mixed sample of the same size and mixture was used for replication trials.

The six cluster methods were studied using four different sets of MMPI indicator-items for each of the two samples; Set I consisted of the 20 most highly discriminant items, Set II consisted of the 30 most highly discriminant items, Set III consisted of the 50 most highly discriminant items, and Set IV of those items in Set III plus 25 nondiscriminant or "garbage" items. Set IV provides the strongest test of the methods and is closest to the actual situation in much taxometric research.

Three of the methods (single linkage, centroid, and median) did not provide acceptable values for the female or male base rates or an adequate hit rate in classifying as to biological sex for any of the four sets of indicator items. In each case the absolute error in the estimate of the male or female base rate exceeded .15, and/or the overall correct classification rate was less than .65 where .50 is the chance rate. In contrast, the results of the other three methods (complete linkage, average linkage, and minimum variance) were much better. For the latter three methods the estimate of female or male base rate was between .40 and .60, and the correct classification rate exceeded .75 for 13 of a total of 24 trial samples.

Whether one views such results as encouraging or otherwise will depend on the research aim and (like it or not) one's philosophy of science. The "professional taxometrician," a typical member of the Classification Society, might be pleased to find three of the popular methods doing as well as they did here; or he or she might not care one way or the other, especially if he or she is philosophically a conventionalist or fictionist, for whom there are no "right answers" to a taxonomic problem.

A physician, psychopathologist, or behavior geneticist should, we think, be less than enthusiastic about findings such as these; and so such workers have been, almost uniformly. Only a tiny minority, surely less than 1%, of clinical psychologists or psychiatrists rely on formal cluster algorithms for help in solving their problems, whether theoretical or clinical. Should we expect them to? Suppose a psychoclinician employed by court services wants to find out whether there are different "types" of sex offenders; how could he or she rationally decide which of the six most popular cluster methods to use in answering such a general question?

Three of them failed to achieve success, by very tolerant criteria, in such a simple, easy task as "telling the girls from the boys." They couldn't manage that on *any* of their eight trials (Golden & Meehl 1981, Table 3). Even the three most accurate were unimpressive as scientific detection instruments would normally be judged. Thus the very best method (average linkage) yielded tolerable results in six trials but failed in two. Would we want to trust such methods in taxonomizing criminals, psychotic patients, school behavior problems, or genetic mental deficiencies? Remember, this is not six out of eight "hits" in diagnosing *individuals*—a validity quite useful for some purposes. Rather the best among the six favorite cluster methods gave only six in eight accurate *nomothetic* (generic, conceptual) results, in detecting the taxon's existence and estimating its base rate.

No one even moderately familiar with the field will find these poor to mediocre results surprising. It is well known that: (1) different cluster algorithms do not tend, by and large, to agree very well with one another; (2) they sometimes detect the taxa generally recognized by competent scholars in a substantive field and sometimes do not; when not, (3) there are no objective (or even generally accepted) criteria telling us which classification to trust; and (4) there are no persuasive *theoretical* considerations for preferring one method over another.

As Meehl (1979, p. 567) points out, it is a striking historical fact that *not one single entity* in psychopathology, neurology, or medical genetics owes its initial discovery to formal cluster analytic methods. He conjectures eight reasons (pp. 571-572) why we might expect this to be so, for which the reader is referred to that paper. See also the critical discussion by the statistician Cormack (1971), and the excellent expositions by Blashfield (1976, 1978), Blashfield and Aldenderfer (1978), and Skinner (1981).

We have not wished to belabor the point that formal cluster algorithms, or at least the received ones in favorite use, have yet to prove their value for the psychoclinician's tasks. It is our own conjecture, which we venture to suggest is rather well corroborated by the evidence to date, that what we have labeled above the "Type II" taxometric problem does not admit of a general solution. We think that there are strong epistemological, mathematical, and domain-substantive (structural, causal) reasons why this is so. That is our justification (other than personal interest and the *Handbook* editors' permission) for focusing this chapter mainly on methodological clarification of the taxometric problem, illustrating that by exposition of our own methods, these latter being *solutions*—strong and test-able ones, as we believe—to taxometric problems of Type III: conjectured latent taxon. It is our conviction that this third kind of problem is by far the commonest, despite its almost complete neglect by taxometricians. So we do not apologize for emphasizing strong Popperian methods in preference to methods of doubtful empirical utility and lacking in coherent theoretical rationale. Readers who remain optimistic about Type II cluster methods can readily find them explained in the excellent standard treatises by Hartigan (1975) and Sneath and Sokal (1973), or the brief summary exposition of Sokal (1974).

THE TAXOMETRIC MODELS

It is desirable to develop a taxometric method so that we know the conditions for

which it will work perfectly. Then we can infer how much the actual latent situation departs from these ideal conditions or "assumptions" of the method and check that this departure is not so great that we are likely to be misled by the parameter estimates. If we do happen to conclude that the departure is too great**,** we can choose to disregard the particular results. The key assumption we used in each of the models described below is that indicators are uncorrelated within the taxon and within the complement of the taxon, referred to below as the "taxon complement" or simply the "complement."

We hypothesize that each individual is either a member of the taxon or a member of the taxon complement. In other literature on taxometric models the taxon and the taxon complement are referred to as "latent classes" (e.g., Clogg, 1977; Goodman, 1975; Green, 1952; Lazarsfeld & Henry, 1968; and Torgerson, 1958). We have emphasized that the taxonomic class should comprise an actual *empirical* taxon, that the classification "carves nature at its joints." Since our purpose is to detect a real, empirical taxonicity, it is not sufficient, for example, merely to produce classifications that optimize a statistical property such as a measure of intraclass homogeneity. We view such a procedure as desirable only to the extent that it helps in detecting real, empirical taxonomies (Meehl, 1979). Because of our emphasis on the empirical nature of the conjectured taxonomic classes, we refer to them as the "taxon" and the "taxon complement" or in general as "taxonomic classes" (for further discussion see Meehl, 1979).

We begin the description of the taxometric models by considering an ideal taxonomic latent situation. There exists a taxonomic class or taxon of individuals with the disease, disorder, or syndrome (denoted by subscript t) and a complementary taxonomic class or taxon complement of individuals without the disorder (denoted by subscript c). Let P be the base rate of the taxon and $Q = 1 - P$ be that of the complement. Suppose we have several dichotomous indicators that discriminate between these two taxonomic classes.

Let p_{tk} be the probability that an individual in the taxon has a positive response to the kth indicator (the "valid positive" rate) and p_{ck} be that for an individual in the complement (the "false positive" rate). Responses are scored as 1 if positive (i.e., deviant, pathological, indicative of the disorder), otherwise as 0. The direction of scoring is determined a priori by empirical or theoretical considerations, but checked empirically by the taxometric method.

Each of the taxometric models we have developed can be derived from the auxiliary conjecture that each pair of indicators (or in the case of the normal model, each pair of components of the indicator, such as items comprising an indicator-scale) are independent within the taxon and within the taxon complement. It is by elimination of the intrataxonomic class covariances (correlations between indicators) that each model becomes overdetermined so we can solve for unique values of the remaining latent parameters such as the taxonomic class base rates and the indicator valid and false positive rates. We trust these parameter estimates only if the auxiliary conjecture specifying independence between indicators within each taxometric class is inferred to be sufficiently close to the actual situation. How we make such inferences is discussed in detail below

CLASSIFICATION OF INDIVIDUALS BY BAYES' THEOREM

When estimates of the latent parameters have been obtained taxometrically, one can then classify individuals, at least fallibly. We use Bayes' Theorem to obtain for each individual an estimate of the model-based probability that the individual is in the taxon. Suppose that a particular individual has a set of indicator scores $(\tilde{x}) = (x_1, x_2, x_3, \ldots x_m)$. If each of these scores were = 1 (code for a positive response), then by Bayes' Theorem the probability of being in the taxon is

$$\frac{P\prod_{i=1}^{m} p_{ti}}{P\prod_{i=1}^{m} p_{ti} + Q\prod_{i=1}^{m} p_{ci}} \tag{6}$$

We can use the model-based estimates of the parameters, P, p_{ti}, and p_{ci} in Bayes' Theorem to obtain an estimate of the probability of an individual being in the taxon. For any set of indicator values (a vector of ones and zeros) the probability of being in the taxon can be obtained from

$$p(t|\tilde{x}) = \frac{Pp(\tilde{x}|t)}{Pp(\tilde{x}|t) + Qp(\tilde{x}|c)} \tag{7}$$

where

$$p(\tilde{x}\,|\,t) = \prod_{i=1}^{m} p_{ti}^{x_i} q_{ti}^{1-x_i} \ ,$$

$$p(\tilde{x}\,|\,c) = \prod_{i=1}^{m} p_{ci}^{x_i} q_{ci}^{1-x_i}$$

$q_{ti} = 1 - p_{ti}$ and $q_{ci} = 1 - p_{ci}$ and we use estimated parameter values. For each individual we can calculate $p(t|\tilde{x})$ and then apply the following classification rule: If $p(t|\tilde{x}) \geq a$, classify as a taxon member; if $p(t|\tilde{x}) < a$, classify as a complement member. It can be shown that the cutting score a is set at .50 to maximize the total number of correct classifications. The use of Bayes' Theorem also requires the assumption that the indicators are independent within each taxonomic class. The error in this "probability score" caused by assumption departure depends on the robustness of the models, a matter that will be discussed below.

AN HEURISTIC METHOD

Indicators that discriminate between the two taxonomic classes and are approximately uncorrelated within each should also behave "consistently" with one another in various ways. For example, if different indicators can be used to obtain multiple estimates of some latent parameter, then these estimates should be sufficiently similar to one another. If the estimates are inconsistent, we know that (for sufficiently large samples) it is likely that one or more of the indicators do not conform to the assumptions underlying the model. Such internal consistency criteria, when used to check the assumptions of a model, can be regarded as statistical tests, christened "consistency tests" by Meehl (1965, 1968, 1973, 1978).

A series of these consistency tests can be used in a consecutive hurdle fashion; when an indicator fails a test, it is removed and not used in the remaining calculations. This series of tests comprises a taxometric method that we call the "consistency hurdles" method. The method is iterative, in that tests are repeatedly applied to the indicators, those failing a test being removed until none of the indicators remaining fail any of the tests. The computational procedures and derivations are partially described below. More mathematical details of the derivations are provided in Golden, Tyan, and Meehl (1974a), Golden and Meehl (1979), and in Meehl (1968). Here we describe in detail the first of four such tests comprising the consistency hurdles method.

Let $y_i = (i = 1, 2, \ldots n)$ denote a dichotomous indicator scored 1 (taxon direction) or 0 (taxon complement direction). Let x_i denote the scale or "key" formed by summing the scores from the $n - 1$ other indicators (all except indicator y_i).

$$x_i = \sum_{j \neq 1}^{n} y_j \,, \qquad (j \neq i) \tag{8}$$

For each "output" indicator y_i and associated "input" scale x_i, we can create the function

$$d_i(c) = a_i(c) - b_i(c) \tag{9}$$

where c is a score on x_i, $a_i(c)$ is the mean of the y_i scores for those individuals with scores *above* c on x_i, and $b_i(c)$ is the corresponding mean score for those individuals *below* c.

The maximum value of $d_i(c)$ over all of the c-values (the possible range is 0 to $n - 1$) tells us something about the underlying latent situation: If each pair of indicators in a set are correlated mainly because each indicator discriminates the two taxonomic classes, then a very poor choice of c should in general reduce the statistical tendency of x_i and y_i to go together. For example, if a neurologist were dealing with the clinical diagnosis of meningitis versus nonmeningitis, and two indicators under consideration, temperature elevation and marked pain on anteroflexion of the neck, are clinical signs of meningitis, then the two signs should correlate significantly in a mixed sample of meningitic and nonmeningitic persons. But if the neurologist were so unwise as to have chosen a very low cutting point on these two indicators—such as temperature above 99.0° and any sign or complaint, however slight, of stiff neck or reluctance to flex the neck—then considerable numbers of patients without meningitis but with other milder infectious conditions, including the common cold, would show one or both of these "signs." Two untoward results would be expected; first, the correlation between the two clinical indicators would be reduced, and second, the identification of the taxonomic class of interest (meningitis) would be poorer.

Monte Carlo studies of the d_i functions for a wide variety of artificial data samples have indicated that its maximum value over all values of c must be at least .10 for the taxometric method to give sufficiently accurate results. Since in practice the estimate of this maximum value can be misleading due to sampling error, especially for samples of sizes less than 500, it is best to smooth the d_i curve by a method such as moving averages. Based upon these smoothed values, the first consistency test requires that max $[d_i(c)] \geq .10$. After the deletion of those k

indicators that fail this requirement, the x_i keys are then recalculated, using the remaining set of $n - k - 1$ indicators and the process is repeated until no items are deleted. The remaining hurdles of the method consist of checking that the indicators, when used in the output role, produce consistent estimates of the hitmax cut on the input scale, are sufficiently discriminating between the taxon and complement, and produce sufficiently consistent estimates of the taxonomic class base rates. The indicator positive rates for the taxon and the complement (i.e., the valid positive and valid negative rates, respectively) can be estimated by the use of the tails of the distribution on the input scale. That is, the method provides an estimate of the taxon base rate and the valid and false positive rates for each of the remaining indicators. The above description of the method is very brief; for a more complete description see Golden and Meehl (1979).

If the results of these consistency tests lead to acceptance of the latent model as an adequate approximation of the state of nature, then Bayes' Theorem can be used to calculate the probability of each individual's belonging to the taxon.

An empirical test in which the MMPI items previously described were used to identify the sexes suggests how well the method can work with real data. For example, the method was applied to the total 75 items (the toughest test used in testing the cluster methods) for a sample consisting of 100 males and 100 females. The estimates of the indicator taxonomic class positive rates and the male and female base rate estimates for the nine items selected by the method are presented in Table 5.1. As one can see, even with a sample of this small size the item positive rates were estimated quite accurately, as was the base rate (.54 vs. .50). The results of other such trials of this method are given in Golden and Meehl (1979) along with the results of applying this method to the schizoid taxon problem.

Table 5.1. The Estimates of the Male and Female Indicator Positive Rates and the Female Base Rate for 75-Item Male-Female Trial of the Consistency Hurdles Method ($N = 200$)

	Males			*Females*			*Estimate of female base rate*
Item	*Estimate*	*True value*	*Error*	*Estimate*	*True value*	*Error*	
1	.05	.08	–.03	.66	.64	.02	.53
2	.09	.10	–.01	.37	.30	.07	.57
3	.41	.50	–.09	.87	.90	–.03	.55
4	.42	.44	–.02	.84	.74	.10	.46
5	.46	.40	.06	.91	.84	.07	.48
6	.25	.36	–.11	.87	.78	.09	.54
7	.05	.14	–.09	.54	.52	.02	.56
8	.08	.08	.00	.47	.38	.09	.55
9	.30	.36	–.06	.80	.84	–.04	.57
						Average:	.54
						True value:	.50

THE MAXCOV-HITMAX MODEL

This model requires the use of three or more "graded" indicators rather than several dichotomous indicators. The indicators may be, for example, MMPI "keys" or scales formed by summing 20 or so MMPI items, each scored 1 and 0 as done previously.

We conjecture that we have a set of three or more indicator-scales, each of which has some validity for discriminating between the two taxonomic-classes, but which are uncorrelated with one another within each taxonomic-class. We need not make further assumptions that the frequency distribution for each indicator is unimodal within taxonomic-class, nor do we need further assumptions regarding the distributions, such as normality, symmetry, homogeneity of variance, or the like.

We will use the following notation.

W, X, Y: indicators and associated random variables

N: compound sample size

W_0: an arbitrary but fixed interval of indicator W

P: base rate of the taxon

$Q = 1 - P$: base rate of the complement

$p(w_0)$: proportion of those in the w_0 interval who are in the taxon

$\mu(X)$: compound population mean for indicator X

$\mu_t(X)$: taxon population mean for indicator X

$\mu_c(X)$: complement population mean for indicator X

$\sigma^2(X)$: compound population variance for indicator X

$\sigma_t^2(X)$: taxon population variance for indicator X

$\sigma_c^2(X)$: complement population variance for indicator X

$\sigma(X,Y)$: compound covariance for indicators X and Y

$\sigma_t(X,Y)$: taxon covariance for indicators X and Y

$\sigma_c(X,Y)$: complement covariance for indicators X and Y

$\sigma(X,Y|w_0)$ compound population covariance for indicators X and Y for subpopulation where $W = w_0$

$\sigma_t(X,Y|w_0)$: taxon population covariance for indicators X and Y for subpopulation where $W = w_0$

$\sigma_c(X,Y|w_0)$: complement population covariance for indicators X and Y for subpopulation where $W = w_0$

A carat above any of these parameters will be used to denote an estimate of the parameter.

Let W, X, and Y be three such indicator-scales, so that W is the "input" indicator and X and Y are the "output" indicators. The (latent) taxonomic class distributions on the input indicator are estimated by using manifest relationships between the two output variables.

As before, the input-output terminology is used merely to describe statistical

procedures; nothing about causal relationships is implied. The taxon is the taxonic class with the higher scores on each of the input indicators and the complement class is that with the lower scores.

The covariance between X and Y for any interval w_0 of indicator W is

$$\sigma(X,Y \mid w_0) = p(w_0)\sigma_t(X,Y \mid w_0) + q(w_0)\sigma_c(X,Y \mid w_0) + p(w_0)q(w_0)v(X \mid w_0)v(Y \mid w_0) \tag{10}$$

where $p(w_0)$ is the proportion of individuals in w_0 interval that are members of the taxon, $q(w_0)$ is the corresponding complement proportion, so that $p(w_0) + q(w_0) = 1$, $\sigma_t(X,Y|w_0)$ is the latent covariance between X and Y for the taxon in interval w_0, $\sigma_c(X,Y|w_0)$ is the corresponding complement covariance, $v(X|w_0)$ is the mean on X for the taxon less that for the complement in interval w_0, and $v(Y|w_0)$ is the corresponding mean difference on Y. Letter v is chosen for "[crude] latent *validity*."

Earlier in this chapter we showed that a procedure for estimating the latent parameters can be derived from the following assumptions:

$$A_1: \ \sigma_t(X,Y|w_0) = \sigma_c(X,Y|w_0) = 0 \tag{11}$$

for each w_0 interval, and

$$A_2: \ \sigma_t(X,Y) = \sigma_c(X,Y) = 0 \ . \tag{12}$$

We studied by Monte Carlo method the robustness and accuracy of the MAXCOV-HITMAX model with respect to these assumptions. Artificial data were generated to produce multivariate normal distributions within each of the two taxonomic classes, because data of this kind are easily generated and can serve the purpose of an exploratory Monte Carlo study (for method see Golden et al., 1974c). Certain parameters of the latent situation were fixed because of constraints and cost. For each of the artificial data samples three indicators were used in the three different input-output role combinations. The values of the indicator taxonomic class means and standard deviations were assigned the same values for each of the three indicators, and each artificial number was rounded off to the nearest integer. For each set of latent parameter values (taxonic base rates, means, and sigmas on each indicator) 25 independent random samples were generated, and each was analyzed by the MAXCOV-HITMAX method.

To generate an artificial sample we specified: (1) the parameters of the multivariate normal distribution for the population taxon and complement; and (2) the taxonic base rates. The parameters of a multivariate normal distribution are the three indicator means and standard deviations and the $\binom{3}{2} = 3$ covariances or correlations between the indicators taken pairwise (Lindgren, 1962). The MAXCOV-HITMAX method is based on the assumption that these correlations are zero for each taxonomic class, and we can test the method for robustness with respect to this assumption by assigning nonzero values to these population correlations.

The various sets of parameter values are described in Table 5.2 (by columns labeled N, P, μ_c, μ_t, σ_c and σ_t). Twenty-five random samples were generated for

each of the 24 latent conditions. In summarizing the results of applying the MAXCOV method to the 24 × 25 = 600 random samples, parameter estimates were regarded as accurate enough for our research purposes if the base rate and hit rate estimates were within .10 of the true (latent) values, and if the taxonomic class mean and sigma estimates were within one interval (usually about ½ of a taxonomic class standard deviation) of the true values.

Table 5.2. Population Parameter Values for the Monte Carlo Trials of the MAXCOV-HITMAX Method

Set	*Variable*	*N*	*P*	μ_c	μ_t	σ_c	σ_t	*v*	σ_t/σ_c	ρ		*#F*
1.1	*N*	1000	.5	8	12	2	2	2	1	0	*	0
1.2		800	.5	8	12	2	2	2	1	0	*	0
1.3		600	.5	8	12	2	2	2	1	0	*	0
1.4		400	.5	8	12	2	2	2	1	0	*	0
2.1	*P*	1000	.6	8	12	2	2	2	1	0	*	3
2.2		1000	.7	8	12	2	2	2	1	0	*	2
2.3		1000	.8	8	12	2	2	2	1	0	*	8
2.4		1000	.9	8	12	2	2	2	1	0		0
3.1	*v*	1000	.5	9	12	2	2	1.5	1	0	*	0
3.2		1000	.5	10	12	2	2	1	1	0	*	15
3.3		1000	.5	11	12	2	2	0.5	1	0		0
3.4		1000	.5	12	12	2	2	0	1	0		0
4.1	σ_t/σ_c	1000	.5	8	12	1.9	2.1	2	1.1	0	*	0
4.2		1000	.5	8	12	1.7	2.3	2	1.3	0	*	0
4.3		1000	.5	8	12	1.5	2.5	2	1.7	0	*	0
4.4		1000	.5	8	12	1	3	2	3	0		0
5.1	*p*	1000	.5	8	12	2	2	2	1	.1	*	0
5.2		1000	.5	8	12	2	2	2	1	.3	*	0
5.3		1000	.5	8	12	2	2	2	1	.5	*	8
5.4		1000	.5	8	12	2	2	2	1	.8	*	0
										ρ_c/ρ_t		
6.1	*N*	1000	.8	8	12	2	2	2	1	.5/.125		0
6.2	$\rho_c/\rho_t = 4$	800	.8	8	12	2	2	2	1	.5/.125		0
6.3		600	.8	8	12	2	2	2	1	.5/.125		0
6.4		400	.8	8	12	2	2	2	1	.5/.125		0

KEY
N: sample size
P: base rate of the taxon
μ_t: mean of the taxon on each indicator
μ_c: mean of the taxon complement on each indicator
σ_t: standard deviation of the taxon on each indicator
σ_c: standard deviation of the complement on each indicator
v: $(\mu_t - \mu_c)/\sigma$ where $\sigma = (\sigma_t + \sigma_c)/2$
ρ: correlation between indicators within the taxon and the complement
*: parameter estimates are always or nearly always accurate
#F: number of failures of consistency tests in 25 samples

We now describe the results. First, different sample sizes (N) of 1000, 800, 600, 400, for $\sigma_c^2 = \sigma_t^2 = \sigma^2$, a difference between the taxonomic class means of 2σ, $P = .5$, and zero taxonomic class correlations, each gave average errors of .01 (2%) in the estimation of P, and less than $\sigma/4$ (½ of an indicator interval) in the estimation of the taxonomic class means and standard deviations. Second, different base rates of .6, .7, .8, and .9 for $N = 1000$, $\sigma_t = \sigma_c$, taxonomic class mean separations of 2σ, and zero taxonomic class correlations, gave corresponding average errors of .03, .04, .02, and .60 in the estimation of the base rate and average errors of less than 3 $\sigma/8$, $\sigma/2$, σ, and $3\sigma/2$ in the estimation of the taxonomic class means and standard deviations. Third, different taxonomic class mean separations of $3\sigma/2$, σ, and $\sigma/2$, for $N = 1000$, $\sigma_t = \sigma_c$, $P = .5$ and zero taxonomic class correlations, gave average errors of .01 in the estimation of P and less than $\sigma/4$ in the estimation of the taxonomic class means and standard deviations. Fourth, different standard deviation ratios (σ_t / σ_c) of 11/10, 4/3, 5/3, and 3 for $N = 1000$, taxonomic class mean separations of $\frac{1}{2}(\sigma_t + \sigma_c)$, $P = .5$ and zero taxonomic class correlations, gave average errors of .02, .03, .08, and .14 in the estimation of P and average errors less than $\sigma/4$, $\sigma/4$, $\sigma/4$, $\sigma/2$ in the estimation of the taxonomic class means and standard deviations. Fifth, different taxonomic class correlations of .1, .3, .5. and .8 for $N = 1000$, taxonomic class mean separations of 2σ, $\sigma_t = \sigma_c = \sigma$, and $P = .5$ gave average errors of .01 in the estimation of P and $\sigma/4$, $\sigma/4$, $\sigma/2$, and σ, in the estimation of the taxonomic class means and standard deviations.

To summarize, the MAXCOV-HITMAX model requires the following conditions in order to provide base rates accurate to within .10 and taxonomic class means and standard deviations accurate to within $\sigma/2$ (an indicator interval):

1. Base rates not disproportionate more than (.2/.8).
2. Separation of means $\geq \sigma$.
3. Standard deviation ratio < 1.7.
4. Taxonomic class correlations ≤ .5.
5. The difference between the two corresponding taxonomic class correlations < .4.

We developed four consistency tests for checking the auxiliary conjectures of the MAXCOV model. This is the analytical basis of one of these tests:

The covariance mixture formula when applied to the taxon and the complement is

$$\sigma(X, Y) = P\,\sigma_t(X, Y) + Q\,\sigma_c(X, Y) + PQ\,v(X)\,v(Y) \qquad (13)$$

where $\sigma(X, Y)$ = manifest covariance for the mixed or compound population

$\sigma_t(X, Y)$ = latent covariance for the taxon

$\sigma_c(X, Y)$ = latent covariance for the complement

P, Q = base rates of taxon and complement

$v(X)$ = difference in means of taxon and complement on indicator x

$v(Y)$ = difference in means of taxon and complement on indicator y

If the assumptions of the model were perfectly met, then $\sigma_t(X,Y) = \sigma_c(X,Y) = 0$ and $\sigma(X,Y) - PQ\, v(X)\, v(Y) = 0$.

We next confront the question "What are the tolerance limits for the quantity $\hat{\sigma}(X,Y) - \hat{P}\hat{Q}\, \hat{v}(X)\, \hat{v}(Y)$?" The carat denotes a model-based parameter estimate. The parameter estimates are erroneous because of sampling error and departure from the model assumptions.

Consider the population parameter

$$T = \sigma(X,Y) - PQ\, v(X)\, v(Y) \tag{14}$$

which can be also written as

$$T = P\sigma_t(X,Y) + Q\sigma_c(X,Y) \tag{15}$$

Since the taxonomic class covariances are generally positive, T is generally positive. It will be useful to consider the differential of T, which is

$$dT = \frac{\partial T}{\partial \sigma(x,y)} d\sigma(x,y) + \frac{\partial T}{\partial P} dP + \frac{\partial T}{\partial v(x)} dv(x) + \frac{\partial T}{\partial v(y)} dv(y) \tag{16}$$

We can interpret this result by considering $d\sigma(X,Y)$ as the error in $\hat{\sigma}(X,Y)$ due to sampling, dP as the error in $\hat{P}$ due to sampling and assumption departure, and likewise for $dv(X)$ and $dv(Y)$. It follows then that $dT = \hat{T} - T$ is the resulting or propagated error in T caused by those errors in $\hat{\sigma}(X,Y)$, $\hat{P}$, $\hat{v}(X)$ and $\tilde{v}(Y)$. The above equation for the differential of T is approximately true for small errors, the approximation being better the smaller the errors. The partial derivatives obtained by differentiation of the above covariance mixture equation are

$$\frac{\partial T}{\partial \sigma(X,Y)} = 1 \tag{17}$$

$$\frac{\partial T}{\partial P} = (2P-1)v(X)v(Y) \tag{18}$$

$$\frac{\partial T}{\partial v(X)} = -PQ\, v(Y) \tag{19}$$

$$\frac{\partial T}{\partial v(Y)} = -PQ\, v(X) \tag{20}$$

Substituting parameter estimates into the expressions for these partial derivatives gives

$$\begin{aligned}\hat{T} \simeq\ & \hat{\sigma}(X,Y)-\sigma(X,Y)+ \\ & (2\hat{P}-1)\hat{v}(X)\hat{v}(Y)(\hat{P}-P)- \\ & \hat{P}\hat{Q}\hat{v}(Y)\big(\hat{v}(X)-v(X)\big)- \\ & \hat{P}\hat{Q}\hat{v}(X)\big(\hat{v}(Y)-v(Y)\big)\end{aligned} \tag{21}$$

We can proceed by specifying an upper limit for the absolute value of the difference in the estimate and actual value for each of

$$\sigma(X,Y),\ P,\ v(X), \text{ and } v(Y).$$

Specifically, let us require that

$$\begin{aligned}&\text{(a) } |\hat{P}-P| \le .10 \\ &\text{(b) } |\hat{v}(X)-v(X)| \le \tfrac{1}{2}\sigma_a(X) \\ &\text{(c) } |\hat{v}(Y)-v(Y)| \le \tfrac{1}{2}\sigma_a(Y)\end{aligned} \tag{22}$$

(where $\sigma_a(X) = \frac{1}{2}\,[\sigma_t(X) + \sigma_c(X)]$ and $\sigma_a(Y) = \frac{1}{2}\,[\sigma_t(Y) + \sigma_c(Y)]$).

Finally, it is very likely that

$$\text{(d) } |\hat{\sigma}(X,Y)-\sigma(X,Y)| \le \frac{4\sigma(X)\sigma(Y)}{\sqrt{N}} \tag{23}$$

Since the MAXCOV-HITMAX method usually produces taxonomic class standard deviation estimates that tend to be too large, we have

$$\begin{aligned}|\hat{T}| \le\ & \frac{4\hat{\sigma}(X)\hat{\sigma}(Y)}{\sqrt{N}} \\ & +|2P-1|\hat{v}(X)\hat{v}(Y)(.10) \\ & +\hat{P}\hat{Q}\hat{v}(Y)\Big(\tfrac{1}{2}\hat{\sigma}_a(X)\Big) \\ & +\hat{P}\hat{Q}\hat{v}(X)\Big(\tfrac{1}{2}\hat{\sigma}_a(Y)\Big)\end{aligned} \tag{24}$$

as our final result. The consistency test now is to compare the quantity

$$|\hat{T}| = \big|\hat{\sigma}(X,Y)-\hat{P}\hat{Q}\hat{v}(X)\hat{v}(Y)\big| \tag{25}$$

with the limit given in the above inequality. If the inequality is satisfied, the test is passed by our data set; if not, it is failed. This consistency test turns out to be a sensitive detector of situations in which the intrataxonomic class correlations are too large and result in inaccurate parameter estimates (sets 5 and 6 of Table 5.2). This sensitivity of the test is reasonable, since its derivation rests squarely on the assumption that these correlations are zero.

In a similar manner we developed three other tests designed to detect other kinds of assumption departure (Golden, Tyan, & Meehl, 1974b).

The most significant result of the Monte Carlo study was that every sample that produced inadequate parameter estimates failed at least one of the four consistency tests and was therefore detectable as untrustworthy (see Table 5.3). Only in set 3.2, where the separation between the taxonic means was small (σ), were the parameter estimates acceptable and incorrectly rejected. But these samples produced only marginally acceptable parameter estimates. So the consistency tests worked nearly perfectly for these artificial data samples. The few instances (6%) in which the consistency tests failed were "conservative" in that accurate results were needlessly rejected.

Table 5.3. Summary of Consistency Test Results for Monte Carlo Study of the MAXCOV-HITMAX Method

		Consistency test "advice"		
		Accept sample	*Reject sample*	
Sample's actual properties	Accurate	336	36	372
	Inaccurate	0	228	228
		336	264	600

Proportion of samples that were correctly accepted or rejected = .94

EMPIRICAL TRIAL OF THE MAXCOV-HITMAX MODEL

The MAXCOV-HITMAX method was also subjected to an empirical trial of detecting the biological sexes with MMPI keys (Golden et al., 1973). Three keys of 20 items each were constructed from the 60 most discriminating items (no "garbage" items were included) selected as described previously.

The actual and estimated male and female taxa indicator distributions and the corresponding descriptive statistics for each of the three arrangements of the indicators are given in Table 5.4. That the parameter estimates are accurate enough for most research in the area of psychopathology measurement is clear simply by inspection. Usual tests of significance for comparing the actual and estimated frequency distributions are of no interest.

Inspection of Table 5.4 also shows the actual (sample) hitmax intervals for keys 1, 2, and 3 to be 12, 9, and 10 respectively. The corresponding maximum covariances occur in intervals 12, 9, and 10, which are in perfect agreement with the true sample values.

The predicted sex can be compared with the actual sex in terms of a hits-misses table. The proportion of "female" predictions that were correct was .90, the corresponding "male" hit rate was .81, and the overall hit rate was .86. Such a hit rate must surely be pushing the maximum theoretically possible with the MMPI item pool.

Table 5.4. The Actual and MAXCOV Estimated Taxa Frequency Distributions for Each Key (N = 1105)

	Key 1				*Key 2*				*Key 3*			
	Male		*Female*		*Male*		*Female*		*Male*		*Female*	
Score	*Actual*	*Estimated*	*Actual*	*Estimated*	*Actual*	*Estimated*	*Actual*	*Estimated*	*Actual*	*Estimated*	*Actual*	*Estimated*
1	0	0	0	0	1	1	0	0	0	0	0	0
2	0	0	0	0	7	7	0	0	4	4	0	0
3	3	3	0	0	13	9	0	4	3	3	0	0
4	11	10	0	1	33	26	3	10	12	11	0	1
5	8	7	0	1	48	36	3	15	26	24	2	4
6	32	34	3	1	63	44	13	32	54	53	3	4
7	41	39	5	7	67	61	17	23	53	60	8	1
8	50	46	9	13	62	59	27	30	57	64	16	9
9	58	66	20	12	59	59	59	59	71	92	27	6
10	60	71	30	19	36	29	87	94	46	58	69	57
11	65	76	27	16	22	52	110	80	43	13	83	113
12	53	58	62	57	11	26	97	82	31	22	91	100
13	29	8	93	114	6	32	91	65	17	14	93	96
14	10	15	110	105	1	18	83	66	9	23	109	95
15	7	18	103	92	0	5	51	46	4	4	71	71
16	2	13	107	96	1	10	26	17	0	3	56	53
17	1	3	61	59	0	1	5	4	0	0	27	27
18	0	2	28	26	0	0	2	2	0	0	15	15
19	0	2	14	12	0	0	1	1	0	0	5	5
20	0	0	3	3	0	0	0	0	0	0	0	0
Base rate	.389	.426	.611	.574	.389	.430	.611	.570	.389	.405	.611	.595
Mean	9.57	9 94	14.10	14.12	7.31	8.66	11.68	10.97	8.60	8.61	12.84	12.95
SD	2.55	2.84	2.56	2.50	2.42	3.17	2.47	2.94	2.57	2.62	2.53	2.40

In order to evaluate the accuracy of the prediction of biological sex by the MAXCOV-HITMAX model, several methods of prediction of biological sex as a *dependent* (*criterion*) *variable* with the indicator variables as the independent variables were tried. None of these methods, including Fisher's linear discriminant function, did significantly better than the MAXCOV method in that the overall hit rate never exceeded .91.

THE NORMAL MODEL

We now develop a taxometric model based on the auxiliary conjecture that the indicator distributions are normal within each taxonomic class. When the MAX-COV-HITMAX and the normal models were first developed, little consideration was given to the relative merits of two basic assumptions, those of zero overall and conditional covariance within taxonomic class and of normality within taxonomic class. The different starting points were chosen mainly because of the resulting mathematical tractability. However, it turns out, as we shall see, that the two models can be developed from the same general assumption. When indicator keys consist of summed MMPI items, it is sometimes reasonable to assume normality within taxonomic class because of the generalized version of the central limit theorem (Von Mises, 1964, p. 302). This theorem states that the distribution of the sum of many different independent Bernoulli variables with virtually any set of values for their means approximates a normal distribution. It follows then that independence between the items or any kind of dichotomous components of the indicator scales within each of the taxonomic classes is a sufficient assumption for both the normal and MAXCOV-HITMAX models.

Whereas the MAXCOV-HITMAX model discussed in the last section requires that there be only two taxonomic classes, the normal model is easily generalized to any number of taxonomic classes. The model was developed first for use with just one indicator (Meehl, 1968) but it was later generalized for any number of indicators (Golden, Tyan, & Meehl, 1974).

One method of solving for the parameter estimates is by trial and error. The procedure for this kind of numerical solution for the single indicator dichotomous taxonomy case is extremely simple. Considering the three taxonomic class distribution parameters required on the normality assumption (i.e., the base rate, mean and standard deviation), we choose a triplet of these values (P, μ_t, σ_t) for the taxon. If we have a sufficiently large sample, then the manifest sample values determine the corresponding parameters (Q, μ_c, σ_c) for the complement. The procedure therefore consists of assigning arbitrary (sliding) values to the base rates P, Q, ($Q = 1 - P$), then to the latent means, μ_t, μ_c, and finally to the latent standard deviations σ_t, σ_c. This logical tree terminates in predicted resultant values for the observed (mixed) frequency distribution. We then compute a chi-square on the discrepancy between the predicted and observed frequencies. It serves first as a significance test (testing departure from the postulated latent model-cum-parameter values) but also, more importantly, as a rough measure of the poorness of our approximation. See Meehl et al. (1969) and Golden and Meehl (1973a) for empirical trials of this numerical method. Suffice it to say here that the method

generally provided estimates comparable in accuracy to those of the MAXCOV-HITMAX model in the detection of the biological sex taxonomy with MMPI indicator scales.

It has been shown that the maximum-likelihood method and the minimum chi-square method produce the same results for large enough samples (Cramér, 1946). However, the maximum-likelihood method requires much less calculation. An outline of the calculations for the maximum likelihood solution is given below; the interested reader is referred to the original article by Hasselblad (1966) for further analytical development.

Suppose that there are n taxonomic classes denoted by the subscript j with distributions on an indicator X and the taxonomic class means, variances, and base rates denoted by μ_j, σ_j, and p_j ($\sum_{j=1}^{n} p_j = 1$). Let X be divided into N intervals denoted by the subscript i so that the interval width is small compared to each σ_j. We assume that the density of the jth taxonomic class in the ith interval, represented by q_{ij}, is approximated by

$$q_{ij} = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[\frac{-(x_i - \mu_j)^2}{2\sigma_j^2}\right] \tag{26}$$

Let Q_i be the compound density for the ith interval so that

$$Q_i = \sum_{j=1}^{n} q_{ij} p_j \tag{27}$$

Hasselblad (1966) has shown that the maximum-likelihood estimates of the unknown latent parameters can be found by the steepest descent iterative procedure. The only required manifest parameter values are the compound sample distribution interval frequencies f_i, $i = 1, 2, 3, \ldots, N$. The procedure results in the following equations:

$$\mu_j = \left[\sum_{i=1}^{N} (f_i | Q_i) q_{ij} x_i\right] \Big/ d_j \tag{28}$$

$$\sigma_j^2 = \left[\sum_{i=1}^{N} (f_i | Q_i)(x_i - \mu_j)^2\right] \Big/ d_j \tag{29}$$

$$p_j = \left[\sum_{i=1}^{N} (f_i q_{ij} p_j | Q_i)\right] \Big/ N \tag{30}$$

where

$$d_j = \sum_{i=1}^{N} (f_i | Q_i) q_{ij}$$

The iterative procedure begins with initial guesses of μ_j, σ_j, and P_j It has not been analytically determined how accurate the initial guesses must be, or whether convergence to the true values will necessarily obtain. However, several empirical trials have been encouraging with regard to both matters. The initial guesses can be

Table 5.5. The Actual and the Normal Model* Estimated Frequency Distributions for Each Key for Male and Female (N = 1105)

	Key 1				*Key 2*				*Key 3*			
	Male		*Female*		*Male*		*Female*		*Male*		*Female*	
Score	*Actual*	*Estimated*	*Actual*	*Estimated*	*Actual*	*Estimated*	*Actual*	*Estimated*	*Actual*	*Estimated*	*Actual*	*Estimated*
1	0	0	0	0	1	3	0	0	0	0	0	0
2	0	0	0	0	7	7	0	0	4	4	0	0
3	3	4	0	0	13	17	0	0	3	7	0	0
4	11	7	0	0	33	32	3	0	12	16	0	0
5	8	16	0	0	48	52	3	1	26	29	2	1
6	32	29	3	0	63	70	13	4	54	44	3	2
7	41	45	5	1	67	78	17	11	53	59	8	6
8	50	57	9	5	62	72	27	25	57	66	16	15
9	58	61	20	13	59	56	59	48	71	64	27	31
10	60	54	30	29	36	35	87	77	46	53	69	53
11	65	39	27	56	22	19	110	102	43	37	83	79
12	53	24	62	89	11	8	97	112	31	22	91	100
13	29	12	93	118	6	3	91	102	17	12	93	100
14	10	5	110	129	1	1	83	77	9	5	109	100
15	7	2	103	118	0	0	51	48	4	2	71	79
16	2	1	107	89	1	0	26	25	0	1	56	53
17	1	0	61	56	0	0	5	11	0	0	27	31
18	0	0	28	29	0	0	2	4	0	0	15	15
19	0	0	14	13	0	0	1	1	0	0	5	6
20	0	0	3	5	0	0	0	0	0	0	0	0
Base rate	.389	.322	.611	.679	.389	.410	.611	.586	.389	.381	.611	.614
Mean	9.57	8.83	14.10	13.99	7.31	7.09	11.68	11.90	8.60	8.24	12.84	12.97
SD	2.55	2.33	2.56	2.29	2.42	2.30	2.47	2.41	2.57	2.48	2.53	2.48

*Minimum chi-square solution used.

Table 5.6. Examples of Empirical Trials of the Normal Model* Using MMPI Keys to Identify the Sexes

	$\hat{P}_m$	$\hat{\mu}_m$	$\hat{\sigma}_m$	$\hat{P}_f$	$\hat{\mu}_f$	$\hat{\sigma}_f$	χ^2
First Key (N = 1105)							
Initial guess	.500	6.00	3.00	.500	13.00	3.00	
No. of interations							
50	.417	9.45	2.50	.582	14.40	2.16	14.89
100	.449	9.66	2.58	.551	14.51	2.12	14.58
200	.483	9.88	2.67	.516	14.64	2.07	14.37
300	.499	9.98	2.71	.501	14.69	2.05	14.33
True sample value	.389	9.57	2.55	.611	14.10	2.56	
Error	.110	.41	.16	–.110	.59	–.51	
Second Key (N = 1105)							
Initial guess	.432	7.36	2.37	.568	11.97	2.26	
No. of iterations							
10	.428	7.33	2.36	.572	11.95	2.27	13.34
50	.411	7.24	2.32	.589	11.89	2.29	13.28
100	.368	6.98	2.23	.632	11.72	2.35	13.18
True sample value	.389	7.31	2.42	.611	11.68	2.47	
Error	–.021	–.33	–.19	.021	.04	–.12	
Third Key (N = 1105)							
Initial guess	.500	6.00	3.00	.500	13.00	3.00	
No. of interations							
10	.418	8.64	2.54	.582	13.02	2.42	15.25
50	.418	8.59	2.48	.582	13.06	2.38	14.92
100	.416	8.58	2.48	.584	13.06	2.38	14.92
200	.409	8.54	2.46	.591	13.03	2.40	14.91
True sample value	.389	8.60	2.57	.611	12.84	2.53	
Error	.020	–.06	–.11	–.020	.19	–.13	

*Maximum-likelihood solution used.

obtained by a method that makes use of probability paper (Harding, 1949), but subsequent study has indicated that reasonable estimates will suffice for most detectable taxonic situations. Each of the three MMPI keys used for the trial of the MAXCOV model was analyzed. The results are given in Tables 5.5 and 5.6. It is seen that the single indicator normal method gave accurate parameter estimates on the second and third keys after just 100 iterations, and it was about at this point that the series of estimates showed convergence. However, for the first key, the base rate estimate is only marginally acceptable, especially after the process had been continued until convergence was apparent. The exact significance of the larger number of iterations required for apparent convergence is not known; however, the result is illustrative of a general finding from trials of the method that when several hundred iterations are required for convergence, it is an indication

the results may not be accurate. We see from Table 5.6 that the chi-square poorness-of-fit values (comparing the estimated and the observed compound sample frequency distributions) do not approach significance for the three keys.

An obvious consistency test for the normal model compares the estimated compound distribution with the observed. The chi-square measures discrepancies between frequencies and is a statistic providing a significance test. But here we do not ask whether the chi-square exceeds a critical value for statistical significance. Rather we ask whether it exceeds some "practical" value that casts doubt that accurate parameter estimation and nonspurious taxon detection obtain We have found that such a critical value for the chi-square parameter does not seem to exist.

Many of the consistency tests suggested for the MAXCOV-HITMAX model can be used with the normal model. Several other tests flowing directly from the normality assumption are possible.

One such test rests on the fact that the sum of two or more normally distributed variables is also normally distributed. Suppose we have two indicators X and Y and use the method with each and with a third indicator formed from the sum of the indicators. The base rate estimates should be the same when the indicators are used singly as when used as a sum. The mean and variance of the sum of indicators within each taxonomic class are given by

$$\mu_t(X+Y) = \mu_t(X) + \mu_t(Y) \tag{31}$$

$$\mu_c(X+Y) = \mu_c(X) + \mu_c(Y) \tag{32}$$

$$\sigma_t^2(X+Y) = \sigma_t^2(X) + \sigma_t^2(Y) + 2\sigma_t(X,Y) \tag{33}$$

and

$$\sigma_c^2(X+Y) = \sigma_c^2(X) + \sigma_c^2(Y) + 2\sigma_c(X,Y) \tag{34}$$

where $\sigma(X,Y)$ is the intrataxonic covariance for the pair of indicator keys. We see then that when two indicators are summed, the single intrataxonic-class covariance between the two indicators can be estimated from the above equation. If both indicators are made up of, for example, MMPI items randomly assigned to each, then this covariance estimate should be sufficiently close to zero.

THE BOOTSTRAP PROCESS

We have repeatedly urged that successively improving the agreement of the model to the data should be one of the fundamental aims of taxometrics. If many bootstrap steps or iterations are required, care must be taken that excessive "psychometric drift" (Loevinger, 1957) does not occur. Through many iterations, the shared content of the set of indicators may change substantially, especially if only a few are found to fit the assumptions of the model. Consistency tests cannot ensure against "convergence" in a pseudo-bootstrap sequence resulting in erroneous inferences. It may seem that one has successfully used a bootstraps procedure when in fact there is decreased verisimilitude due to psychometric drift. The convergence is to the wrong latent situation, or at least to one further from the correct latent situation than obtained at an earlier stage of the process. Since this conceptual danger exists in taxometric bootstrapsing, as in *all* empirical inference

to latent causal entities, we look for ways to corroborate the desired increase in verisimilitude. We have used the following multiple criteria:

1. Select indicators with clinical or face validity.
2. Further select those that are positively correlated in the mixed sample.
3. Delete or combine indicators that are inferred to be highly correlated within a taxonomic class. If a pair of indicators is highly correlated in this way, then one of the indicators should be deleted, or possibly, if it is substantively desirable, the two indicators should be combined to form a single indicator.
4. Delete the indicators that do not discriminate sufficiently between the taxonomic classes.
5. Check the parameters for reasonableness with regard to considerations such as the following:
   a. Rank order of indicators in terms of validity.
   b. Size of *P* compared to previous research and to clinical experience.
   c. Indicators with a low false positive rate.
   d. Indicators with a high valid positive rate.
6. Repeat the analysis with the modified indicator set if criteria (3), (4), or (5) are not met.

DETECTION OF THE SCHIZOID TAXON

In a pilot study (Golden & Meehl, 1979) MMPI items were selected as candidate indicators of schizoidia. Items were required to discriminate between schizophrenics and normals, the hypothesis being that such items have better average potential to discriminate between schizoids and nonschizoids than items nonvalid against formal diagnosis. Fifty-three items discriminated between 96 diagnosed schizophrenics and the Minnesota normal sample by a difference of at least .20.

Next we required that the item not be highly correlated with decompensation-related variables, such as severity of illness. This requirement should reduce selection of items highly correlated with each other within the schizoid group mainly through underlying decompensation-related variables. We also wanted items that do not discriminate appreciably among other diagnostic classes (e.g., psychotics vs. neurotics), but were able only to require that an item not discriminate highly among diagnosed subtypes of schizophrenia, or among those other psychoses for which we had samples of sufficient size. These "negative" requirements, aimed to minimize nuisance covariance (Meehl 1972b, pp. 160-174), were failed by 20 of the 53 previously selected items, leaving 33 items for further analysis.

The sample used consisted of 211 male inpatients at the University of Minnesota Hospital who had been diagnosed as having a neurosis, personality disorder, or transient situational disorder. No diagnosed schizophrenics, patients with other psychotic diagnoses, or brain syndromes were included in this sample, so as to reduce the probability of detecting taxonomic classes other than the one of interest.

When the consistency hurdles method was applied to this sample, using the 33 selected MMPI items, it deleted all but seven items. The taxon base rate was estimated to be .37. Applying Bayes' Theorem to the estimated valid and false positive rates and base rate, individuals were classified as either schizoid or not. The Bayes' probabilities tended to be close to zero or one, a result that previous Monte Carlo analyses have shown to indicate a real taxonomy.

If one method correctly classifies a proportion p_1 (a quantity we don't know in practice) of the total mixed sample, while a second method does so with a proportion p_2, and if the two methods make *independent* errors of classification, then the proportion of classifications for which the two methods agree is $p_1p_2 + (1 - p_1)(1 - p_2)$. For example, if $p_1 = p_2 = .80$, then the agreement rate would be .68. An estimate of the correct classification rate, p, for a single set of indicators can be obtained by a method developed in Golden and Meehl (1974, 1979). For each individual, we use Bayes' Theorem to calculate the probability of being schizoid and that of being nonschizoid, and select the larger of these two values. It can be shown that an estimate of the overall correct classification is the average of the larger of these two probability values across all individuals. The model-based, estimated correct classification rate for each of the two sets of items was found to be about .85, which is quite high. The estimated agreement rate between the two classifications should then be about $(.85 \times .85) + (.15 \times .15) = .75$, close to the observed value of .70; so the concordance shows satisfactory agreement with what the separate classification estimates would have predicted.

The MMPI item data were also analyzed by the MAXCOV-HITMAX and normal methods (described earlier). Each of these taxometric methods was used to detect the schizoid taxon, but with different sets of MMPI items than the above.

Three scales were constructed from the 113 items that discriminated between the detected taxonomic classes of "schizoids" and "nonschizoids" by a difference of .20 or more. Factor analysis (Varimax rotation) of the 113 items in the total mixed sample ($N = 211$) suggested three factors; the 20 highest loading items for each factor were selected to form the three scales. When these three keys were used with the MAXCOV method, the taxon detected had a base rate estimate of .40, which agrees well with that of the first method. The MAXCOV method also includes consistency tests that help avoid being misled by inaccurate parameter estimates, and Monte Carlo study indicates that these work quite well (Golden & Meehl, 1973b). The passage of these tests provides additional support that the detected taxon is not spurious.

Factor analysis of the 13 standard MMPI scale scores in the total sample ($N = 211$) produced a Varimax factor that accounted for 41% of the common variance and correlated highly with the Psychasthenia (.69), Schizophrenia (.53), Depression (.61), and Social Introversion (.79) scales; all other loadings were below .30, except for K (–.43). The items in these four scales were combined to make a long scale that was used as the single indicator. The schizoid taxon base rate was estimated by the Normal method to be .41, again in excellent agreement with previous estimates. The chi-square value of 2.3 was nonsignificant and below even the value expected if the assumptions of the method were perfectly satisfied. The difference between the two taxonomic class means on this indicator was

estimated to be about two class standard deviations. Since this is about the same degree of separation obtained for the sexes in the previous analysis using MMPI items, we have additional evidence that the schizoid taxonomy appears to be detectable with MMPI items as indicators. Recent unpublished research indicates that the seven items described above lack validity in other samples; we advise other researchers to use the indicator consisting of the above four MMPI scales in future work.

All three of the taxonomic methods classified individuals as either "in" or "out" of the schizoid taxon, with agreement rates between pairs of methods near to that expectable from estimates of their misclassification rates. This result obtained even though the three methods are based on loosely related assumptions, and were used with different MMPI indicators. Also the three base rate estimates were close to each other and to a personal clinical estimate of .40 to .45 by Meehl; the latter was a prerecorded "impressionistic guess" based on some 30 years of outpatient private practice and the use of the Checklist of Schizotypal Signs (Meehl, 1964).

Each individual whose probability of being a member of the schizoid taxon exceeded .50 was classified as "probable schizoid" and others as "probable nonschizoid," thereby forming two subsamples that could be compared on other variables. The mean MMPI profile for the individuals classified as members of the schizoid taxon, presented in Figure 5.2, was nearly identical to the "2-7-8" code type. The mean MMPI profile for those individuals classified as *not* members of the schizoid taxon was considerably lower, and not similar to any standard code type. This subsample is presumably quite heterogeneous, as one would expect.

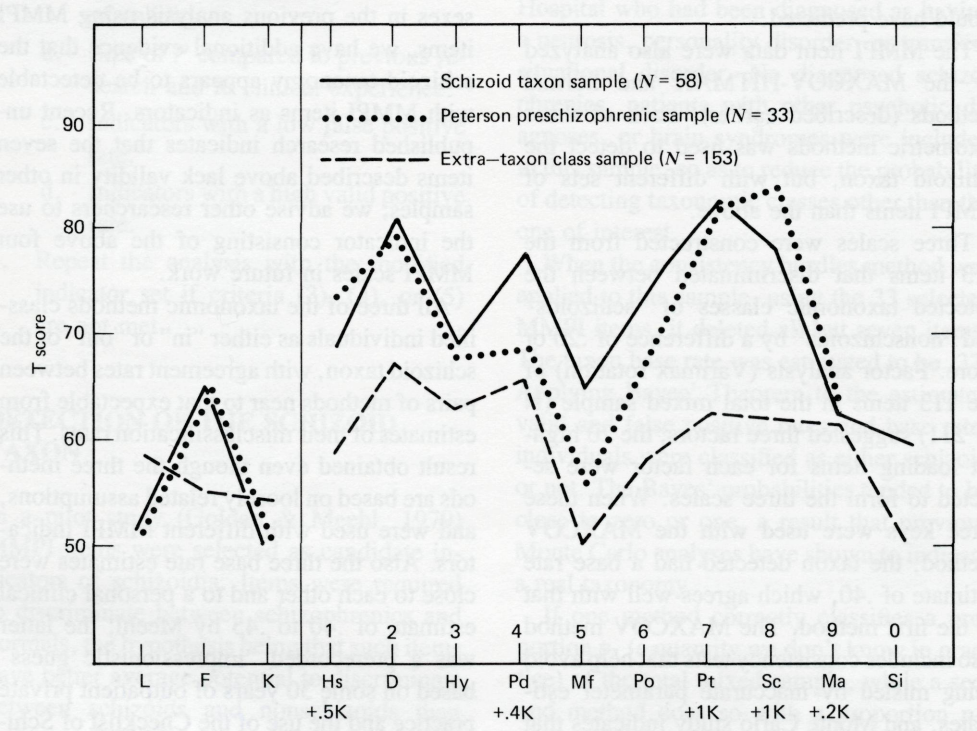

Figure 5.2. Mean NMPI profiles for taxometrically identified schizoids and nonschizoids and for clinically identified preschizophrenics.

The most impressive evidence of construct validity was that the mean MMPI profile of the schizoid taxon was very similar to that of a sample of pre-schizophrenics studied by Peterson (1963). A search of Lanyon's (1968) *Handbook of Group MMPI Profiles* revealed few nonschizoid diagnostic groups that also have the 2-7-8 mean profile pattern exhibited by both the present schizoid taxon and the Peterson preschizophrenic sample.

There probably are neurological and physiological variables more powerful as indicators of schizoidia than MMPI responses. However, MMPI responses, even though far removed causally from any genetic etiological source, may still suffice, in samples of a thousand or more, to provide a convincing test of the existence or nonexistence of the schizoid taxon. The results of the present preliminary trial, even though the sample size is quite small, indicate that such a taxon is likely to exist. These results are sufficiently encouraging to justify an attempted replication of the present study with a much larger sample. The substantive results presented here are not regarded as confirmation of a theory, but are offered in the "context of discovery" (Reichenbach, 1938).

TESTING THE SINGLE DOMINANT GENE HYPOTHESIS

Meehl (1962) hypothesized that schizotaxics are born with a neurological predisposition to acquire a personality organization with four cardinal personality traits: cognitive slippage, ambivalence, anhedonia, and social aversiveness, the inheritance mechanism conjectured to underlie schizotaxia being a dominant gene. This genetic hypothesis is related to the taxonicity aspect of Meehl's theory. A theory taxonic at the genetic level permits stronger refutation tests than does one at the levels of personality or psychopathology.

We have developed an interrelated set of tests that require the use of indicators on the *parents of* schizophrenic probands, who, according to the theory, must carry the schizogene because of their observable phenotypic condition. Suppose each member of the parent-pairs responds (+) or (–) to several dichotomous fallible indicators of the schizogene, such as a personality questionnaire item, a "soft" neurological sign, or whatever. Since the indicator is fallible, the dichotomy on the indicator is imperfectly correlated with the genetic dichotomy. We assume nonassociative mating so that (nearly) always one and only one parent of each pair of parents, under the single dominant gene model, is a schizotype. We assume that in the population of such parents the base rate of schizotypy is exactly one-half.

To make the model mathematically tractable, we make the following auxiliary conjectures. These assumptions are made, for the moment, in the context of discovery, but later will be subjected to indirect empirical testing.

A_1 Each pair of indicators is independent within the schizotypic and within the nonschizotypic populations and

A_2 The response to an indicator by the schizotypic parent is independent of that of the nonschizotypic mate.

Let us use the following notation:

p_s: the proportion of schizotypic parents that respond in + (schizotypic)

direction on a given indicator

p_n: the same for the nonschizotypic parents

A: one of the parents determined by an indexing procedure either dependent or independent of schizotypy (as specified; A will also be used for the class of such parents)

B: the parent (or class of parents) that is not A

C: the (compound) class of all parents (the union of A and B)

p_c: the proportion of the parents that responded in (+) direction

P_{AB}: the proportion of the parent pairs where both parents responded in the (+) direction

σ_{AB}: the covariance between the responses of the A parents with those of B parents for a given indicator

Thus p_s and p_n are latent and unknown, and we first wish to express them as functions of manifest parameters. We have shown that:

Under the assumption A_2 and the single dominant gene hypothesis,

$$p_s = p_c + \left(p_c^2 - p_{AB}\right)^{1/2} \tag{35}$$

$$p_n = p_c - \left(p_c^2 - p_{AB}\right)^{1/2} \tag{36}$$

for any indicator. Derivations of this and the following results are given in Golden and Meehl (1978).

Next we shall propose a set of consistency tests that follow from A_1, A_2, and the single dominant gene hypothesis. If these tests are grossly unsatisfied by the observational facts, a strong suspicion arises that the model represents a poor approximation to the state of nature. Either the estimates of the latent quantities are untrustworthy, or, more importantly, the dominant gene hypothesis is discorroborated. First we have shown that: The covariance between the responses of the A parents with that of the B parents for any such valid indicator must be negative and is equal to the quantity $-\frac{1}{4}(p_s - p_n)^2$ when the indexing is independent of schizotypy.

If we have three indicators X, Y, Z, then under the assumption A_1 we have for indicator X

$$p_s = p_c + 2\left[\frac{\sigma_{xy}\sigma_{xz}}{\sigma_{yz}}\right]^{1/2} \tag{37}$$

and

$$p_n = p_c - 2\left[\frac{\sigma_{xy}\sigma_{xz}}{\sigma_{yz}}\right]^{1/2} \tag{38}$$

where the three covariances refer to those of manifest-compound parent population.

For any indexing independent of schizotypy, we have

$$\sigma_{AB}(X)\frac{-\sigma(X,Y)\sigma(X,Z)}{\sigma(Y,Z)} \tag{39}$$

for indicator X. It should be noted that $\sigma_{AB}(X)$ is the covariance between parent pairs indexed independently of schizotypy, whereas $\sigma(X,Y)$, $\sigma(X,Z)$, and $\sigma(Y,Z)$, refer to the compound parent population between two different indicators.

Further, we have derived the following result: Under the independence assumption A_2 and the single dominant gene hypothesis and for any indexing of the parent pairs, the base rates of schizotypes among the A and B parents are given by

$$P = \frac{1}{2} \pm \frac{1}{2} K \tag{40}$$

where

$$K = \frac{p_A - p_B}{\left[(p_A - p_B)^2 - 4\sigma_{AB}\right]^{1/2}} \qquad \left(P \neq 1/2\right) \tag{41}$$

Any indicator that satisfies A_1, A_2, and the single dominant gene hypothesis will produce the same value of P; in other words, the above expression of manifest parameters for K is invariant across such indicators and thus provides another means of indicator selection.

Once we have determined P and Q as above, we can use the Dawes-Meehl equations (Dawes & Meehl, 1966) to obtain p_s and p_n

If P is determined as in Equation (40), then

$$p_s = \frac{Pp_A - Qp_B}{P - Q} \tag{42}$$

$$p_n = \frac{Qp_B - Pp_A}{P - Q} \tag{43}$$

The two kinds of covariances, those between the two responses of parent pairs to the same indicator X, denoted by $\sigma_{AB}(X)$, and those between two indicators X and Y within either the A or the B parents, denoted by $\sigma_{XY}(W)$, are related according to the next result.

Under the above independence assumptions and the single dominant gene hypothesis,

$$\sigma_{AB}(X)\,\sigma_{AB}(Y) = \sigma_{XY}(W) \tag{44}$$

for any two such indicators X and Y for any indexing of the parents.

SUMMARY

We view taxometrics as the application of formal (mathematical) methods to the problem of detecting nonarbitrary classes (types, species, disease entities, *real syndromes* that exist in the external world—"carving nature at its joints"), with a derivative formal procedure for sorting individuals in or out of such inferred taxa. There are three broad classes of taxometric problems, set by the investigator's state of knowledge and his or her theoretical or clinical aims: In Type I, Accepted

Criterion Prediction, the taxon is known to exist and a defining (or quasi-infallible "proxy") indicator is available in the research context. The aim is to devise, relying on the accepted criterion, a taxometric function of the (more usually available) fallible indicators that will classify future individuals where the accepted criterion is unavailable, at least at the time of classification. In Type II, Classical Cluster Analysis, the existence, number, and nature of taxa in a domain are highly problematic, so the investigator has no accepted criterion and, typically, does not wish to rely on substantive theoretical conjectures even for identifying privileged ("high-validity") indicators. Instead he or she proceeds by constructing a matrix of interindividual similarity measures on a (usually large) number of indicators of unknown relative weight, hoping that by applying a suitable cluster algorithm to this data-summanzing matrix he or she will be able to discern the taxometric structure underlying the observed pattern. In Type III, which we have labeled the Conjectured Latent Taxon Problem, our knowledge situation lies somewhere between that of Type I and Type II, and the taxometric methods we advocate are tailor-made for that intermediate case. From prior knowledge of the domain (experimental studies, statistical analysis of file data, clinical experience, theory, common sense, intuition, or sheer guesswork) we conjecture the existence of a hypothetical taxon.

Our own interest being theoretical, and our philosophy of science being realist rather than instrumentalist or fictionist, we think mainly of "true taxon" as designating an objective entity whose nature will ultimately be understood in some strong *theoretical* (structural, compositional, historical, or causal) sense. The clearest and most interesting examples in psychopathology (and the clinical sciences generally) are those in which the phenotypic taxon arises from a quasi-dichotomous specific etiology (ideally a germ, a major gene, or a traumatic event, but including polygenic or environmental threshold effects, step-functions, Cattell's "environmental molds," etc.). However, while these methodological preferences inform our own thinking and have guided our research, a weaker view of latent taxonicity is compatible with the conjectured latent taxon strategy, since our "taxonicity" concept is pretty well specified implicitly by the formalism and the data.

We think Type III is by far the commonest knowledge/aim situation presented in psychopathology. But whether that is true or not, Type II seems on present evidence, some offered by us here, to have no persuasive general solution, either theoretically or empirically. (See Skinner, 1981 for an illuminating discussion.) Focusing therefore on Type III, the conjectured latent taxon problem, we sketch out the usual mode of informal, nonmathematical discernment of a taxon by clinicians, with special attention to logical and methodological clarification of open concepts, contextual (implicit) definition, alleged circularity of causal explanation via inferred entities, the bootstraps effect, and the role of auxiliary conjectures in empirical testing. We emphasize the necessity for numerical point-estimation or other "risk-taking" predictions, capable of yielding strong Popperian tests of a taxonomic model. We downplay traditional significance testing as a feeble, low-risk way to do science, and we advocate use of multiple joint consistency tests. But strict falsificationism is rejected in favor of specifying

numerical tolerances and approximations, recognizing that all mathematical models in the life sciences have imperfect verisimilitude.[1]

Three new taxometric methods of our devising are explained, the MAXCOV-HITMAX, the consistency hurdles, and the normal. Monte Carlo runs and a biological sex pseudo-problem with real data are offered (although the mathematical derivations in *this* kind of taxometrics "speak for themselves") as evidence of their usefulness. A preliminary study of the nondiagnosed schizoid taxon, identified by MMPI items, is presented. Finally, we derive equations for a theoretical extension of our taxometrics to the problem of testing a dominant gene theory using fallible phenotypic indicators.

REFERENCES

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

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

Bailey, N. T. J. (1965). Probability methods of diagnosis based on small samples. *Mathematics and Computer Science in Biology and Medicine.* HMSA.

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

Blashfield, R. K. (1978). Failure of cluster analysis in psychiatric research. Paper presented at the meeting of the American Psychological Association, Toronto, Canada, August 30.

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

Buchanan, C. E., & Jones, M. B. (1969). A within family study of schizophrenia and a visible subpapillary plexus in the nailfold. *Schizophrenia,* **1**, 61-75.

Carnap, R. (1936, 1937; 1950). Testability and meaning *Philosophy of Science* **3**, 420; **4**, 2. Reprinted with corrigenda and additional bibliography. New Haven: Yale University Graduate Philosophy Club.

Chan, L. S., & Dunn, O. J. (1972). The treatment of missing values in discriminant analysis, I: The sampling experiment. *Journal of the American Statistical Association,* **67**, 473-477.

Clogg, C. C. (1977). *Unrestricted and restricted maximum likelihood latent structure analysis: A manual for users.* (Population Issues Research Office Working paper 1977–04), University Park, Pa.

Cormack, R. M. (1971). A review of classification. *Journal of Royal Statistical Society,* Series A, **134,** 321-367.

Cornfield, J. (1967). Discriminant functions. *Reviews of the International Statistical Institute,* **35**, 142-153.

Cramér, H. (1946). *Mathematical methods of statistics.* Princeton, N.J.: Princeton University Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological*

[1] The senior author, while still skeptical about the atheoretical cluster approach, has recently attempted to generalize our conjectured latent taxon methods to permit a "blind inductive scanning" of miscellaneous indicators for the multiple taxa case (Meehl, 1982). As of this writing, the method has not been tested and a computer program does not yet exist. In the second half of this chapter we briefly describe the analytical and empirical research in taxometrics by the present authors. Recently, Golden has developed an improved bootstraps model (Golden, [1982]) and applied it to the detection of conjectured taxonomic classes of dementia and other disorders in older individuals (Golden et al., in press), schizotypy in children (Golden et al., in press), and kidney disease in children.

Bulletin, **52**, 281-302.

Darlington, R. B. (1968). Multiple regression in research and practice. *Psychological Bulletin.* **69**, 166-168.

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.

Draper, N. R., & Smith, H. (1966). *Applied regression analysis.* New York: Wiley.

Dunn, O. J. (1971). Some expected values for probabilities of correct classification in discriminant analysis. *Technometrics,* **13,** 345-353.

Dunn, O. J., & Varady, P. D. (1966). Probabilities of correct classification in discriminant analysis. *Biometrics,* **22**, 908-924.

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

Finn, J. D. (1974). *A general model for multivariate analysis.* New York: Holt, Rinehart & Winston.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics,* **7**, 179-188.

Freud, S. (1895a, 1962). On the grounds for detaching a particular syndrome from neurasthenia under the description "anxiety neurosis." In J. Strachey (Ed.), *Standard edition of the complete psychological works of Sigmund Freud.* Vol. 3. London: Hogarth Press, pp. 85-115.

Freud, S. (1895b; 1962). A reply to criticisms of my paper on fee anxiety neurosis. In J. Strachey (Ed), *Standard edition of the complete psychological works of Sigmund Freud,* Vol. 3. London: Hogarth Press, pp. 119-139.

Gilbert. E. S. (1969). The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. *Biometrics,* **25**, 505-515.

Goldberg. L. R. (1969). The search for configural relationships in personality assessments: The diagnosis of psychosis vs. neurosis from the MMPI *Multivariate Behavioral Research,* **4**, 523-536.

Golden, R. R. (~~in press~~ 1982). A taxometric model for detection of a conjectured latent taxon. *Multivariate Behavioral Research* **17**, 389-416.

Golden, R. R., & Meehl, P. E. (1973). *Detecting latent clinical taxa. IV: Empirical study of the maximum covariance method and normal minimum chi-square method, using three MMPI keys to identify the sexes* (Rep PR–73–2). Minneapolis: University of Minnesota, Reports from the Research Laboratories of the Department of Psychiatry.

Golden, R. R., & Meehl, P. E. (1973). *Detecting latent clinical taxa, V: A Monte Carlo study of the maximum covariance method and associated consistency tests* (Rep. PR–73–3). Minneapolis: University of Minnesota. Reports from the Research Laboratories of the Department of Psychiatry.

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

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

Golden, R. R., & Meehl, P. E. (1981). Detection of biological sex: An empirical test of cluster methods. *Multivariate Behavioral Research,* **15**, 475-496.

Golden, R. R., & Meehl, P. E. (in preparation). *Taxometric analysis of causal entities: Detection of the schizoid taxon.* New York: Academic Press. [Book not completed]

Golden, R. R., Tyan, S. H., & Meehl, P. E. (1974a). *Detecting latent clinical taxa, VI: Analytical development and empirical trials of the consistency hurdles theory* (Rep PR–74–4). Minnea-

polis: University of Minnesota, Reports from the Research Laboratories of the Department of Psychiatry.

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

Golden, R. R., Tyan, S., & Meehl, P. E. (1974c). *Detecting latent clinical taxa. IX: A Monte Carlo method for testing taxometric theories* (Rep. PR–74–7). Minneapolis: University of Minnesota, Reports from the Research Laboratories of the Department of Psychiatry.

Golden, R. R., Teresi, J. A., & Gurland. B. J. (in press). Detection of taxonomic classes of health and social problems using the Comprehensive Assessment Evaluation and Referral interview schedule. *Journal of Gerontology.*

Golden, R. R., Erlenmeyer-Kimling, L., & Cornblatt, B. (in press). *Taxometric detection of schizotypy in children using a battery of attention, motor and cognitive indicator-tests.* New York: Columbia University. Columbia University Statistical Reports. Division of Biostatistics.

Goldstein, M. (1977). A two group classification procedure for multivariate dichotomous responses. *Journal of Multivariate Behavioral Research,* **12**, 335-346.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association,* **70**, 755-768.

Green, B. F. (1952). Latent structure analysis and its relation to factor analysis. *Journal of the American Statistical Association,* **47**, 71-76.

Grünbaum, A. (1976). Is the method of bold conjectures and attempted refutations *justifiably* the method of science? *The British Journal for the Philosophy of Science*, **27**, 105-136.

Guion, R. M. (1965). *Personnel testing.* New York: McGraw-Hill.

Harding, J. P. (1949). The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of the Marine Biological Association,* **28**, No. 1, 141-153.

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

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

Hays, W. L. (1973). *Statistics for the social sciences.* 2nd ed. New York: Holt, Rinehart & Winston.

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

Kerlinger, F. N., & Pedhazuer, E. J. (1973) *Multiple regression in behavioral research.* New York: Holt, Rinehart & Winston.

Lachenbruch, P. A. (1968). On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlations coefficient. *Biometrics,* **24**, 323-334.

Lachenbruch, P. A. (1975). *Discriminant analysis.* New York: Hafner Press.

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

Lanyon, R. I. (1968). *A handbook of MMPI group profiles* Minneapolis: University of Minnesota Press.

Lazarsfeld, P. F., &. Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Lindgren, B. (1962). *Statistical theory.* New York: Macmillan

Loevinger, J. (1957). Objective tests as instruments of psychological theory *Psychological Reports.* **3**, 635-694.

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

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

Lyle, O. E., & Gottesman, I. I. (1977). Premorbid psychometric indicators of the gene for Huntington's Disease. *Journal of Consulting and Clinical Psychology,* **45**, 1011-1022.

Major, R. H. (1932). *Classic descriptions of disease* Springfield, Ill.: Charles C. Thomas.

Meehl, P. E. (1950). Configural scoring. *Journal of Consulting Psychology,* **14**, 165-171.

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

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

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

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

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.

Meehl, P. E. (1972a). A critical afterword. In I. I. Gottesman & J. Shields, *Schizophrenia and genetics.* New York: Academic Press.

Meehl, P. E. (1972b). Reactions, reflections, projections. In J. N. Butcher (Ed), *Objective personality assessment.* New York: Academic Press.

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

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

Meehl, P E. (1973b). *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. (1979). A funny thing happened to us on the way to the latent entities. *Journal of Personality Assessment,* **43**, 563-581

Meehl, P. E. (1982). *Detecting latent clinical taxa, X: Extension of methods to problem of inductive scanning for multiple taxa* (Rep. No. PR–82–1). Reports from the Research Laboratories of the Department of Psychiatry, University of Minnesota, Minneapolis: University of Minnesota.

[Meehl, P. E. (1990). Toward an integrated theory of schizotaxia, schizotypy, and schizophrenia. *Journal of Personality Disorders*, **4**, 1-99.]

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

Morrison, D. F. (1976). *Multivariate statistical methods.* 2nd ed. New York: McGraw-Hill.

Morrison, D. E., & Henkel, R. E. (Eds.) (1970). *The significance test controversy.* Chicago: Aldine.

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

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Overall, J. E. & Klett, C. J. (1972) *Applied multivariate analyses.* New York: McGraw-Hill.

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

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

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

Peterson, D. R. (1963). The diagnosis of subclinical schizophrenia. In G. S. Welsh & W G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine.* Minneapolis: University of Minnesota Press.

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

Popper, K. R. (1974). Autobiography. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper* LaSalle, Ill.: Open Court.

Rado, S. (1956). *Psychoanalysis of behavior.* New York: Grune & Stratton.

Rado, S. (I960). Theory and therapy: The theory of schizotypal organization and its application to the treatment of decompensate schizotypal behavior. In S. C. Scher & H. R. Davis (Eds), *The outpatient treatment of schizophrenia.* New York: Grune & Stratton, pp. 87-101.

Rado, S., & Daniels, G. (1956). *Changing concepts of psychoanalytic medicine* New York: Grune & Stratton.

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

Rozeboom, W. W. (1966). *Foundations of the theory of prediction.* Homewood, Ill.: Dorsey Press.

Salmon, W. C. (1967). *The foundations of scientific inference.* Pittsburgh: University of Pittsburgh Press.

Skinner, H. A. (1981). Toward the integration of classification theory and methods. *Journal of Abnormal Psychology,* **90**, (1), 68-87

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

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

Solomon, H. (Ed.) (1961). *Studies in item analysis and prediction.* Stanford, Calif.: Stanford University Press.

Tatsuoka, M. M. (1971). *Multivariate analysis: Techniques for educational and psychological research.* New York: Wiley

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley

Tryon, R. C. (1939). *Cluster analysis.* Ann Arbor, Mich.: Edwards Brothers.

Von Mises, R. (1964). *Mathematical theory of probability and statistics.* New York: Academic Press.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin,* **83**, 213-217

Wainer, H. (1978). On the sensitivity of regression and regressors. *Psychological Bulletin,* **85**, 267-273.

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

pdf by LJY, February 2016