Bootstraps Taxometrics

Solving the Classification Problem in Psychopathology

Paul E. Meehl
University of Minnesota

Classification in psychopathology is a problem in applied mathematics; it answers the empirical question "Is the latent structure of these phenotypic indicator correlations taxonic (categories) or nontaxonic (dimensions, factors)?" It is not a matter of convention or preference. Two taxometric procedures, MAMBAC and MAXCOV-HITMAX, provide independent tests of the taxonic conjecture and satisfactorily accurate estimates of the taxon base rate, the latent means, and the valid and false-positive rates achievable by various cuts. The method requires no gold standard criterion, applying crude fallible diagnostic "criteria" only in the phase of discovery to identify plausible candidate indicators. Confidence in the inference to taxonic structure and numerical accuracy of latent values is provided by multiple consistency tests, hence the term coherent cut kinetics *for the general approach. Further revision of diagnostic systems should be based on taxometric analysis rather than on committee decisions based on clinical impressions and nontaxometric research.*

"How shall we classify?" is a scientific or technological question, a problem for applied mathematics. There is a prior epistemological question as to how we evidence a category's reality and a pragmatic question, "Why do we want to classify anyway?" I ask this question not rhetorically but seriously, unlike some dogmatic antinosologists, who wrongly think we know there cannot be any categories of personality or mental disorder. The truth is that we usually do not know whether we are dealing with categories or with dimensions, and in the past we have not had a sound method for finding out the true state of affairs. My interest in developing new taxometric statistics was partly motivated by clinical concerns as a practitioner but was mainly motivated by a theoretical problem: how to test competing genetic models for inheritance of the schizophrenic predisposition—in my theory, *schizotaxia,* a neurological defect that leads to diagnosable schizophrenic illness in only a small fraction of those who have the genotype (Meehl, 1962, 1972, 1989, 1990c, 1990d).

Importance of Valid Categories

Category words are often used without taxonic claim to designate intervals on a quantitative scale or volumes in a descriptor hyperspace, for convenience of communication. It is not clear whether the payoff in ease of communication makes up for the disadvantages (e.g., loss of information). Some examples of such category words are *introverted, bright, dominant, obese,* and *depressed.* The use of such words is particularly dangerous when it is simply assumed that a taxonic entity underlies the category. Even when a category possesses objective existence (as a species, type, or disease entity, i.e., a real taxon), clinical practice may or may not be aided. Grove (1991) showed that over almost all of the parameter space encountered in psychopathology, multiple regression prediction of output variables is superior to taxon-mediated prediction. But that pragmatic surprise does not liquidate our theoretical interest. Philosopher Herbert Feigl (1950) pointed out the research diseconomy of pairwise input-to-output correlations requiring separate empirical study. If one has m input variables (e.g., symptoms, signs, test scores, informant ratings) and n output variables to be predicted (e.g., suicide risk, drug of choice, response to

Editor's note. Articles based on APA award addresses that appear in the *American Psychologist* are scholarly articles by distinguished contributors to the field. As such, they are given special consideration in the *American Psychologist* editorial process.

This article was originally presented as part of a Distinguished Professional Contribution award address at the 102nd Annual Convention of the American Psychological Association in Los Angeles, California, in August 1994.

Samuel M. Turner served as action editor for this article.

Author note. This is an edited version of lectures given on receipt of the Joseph Zubin Award at the meeting of the Society for Research in Psychopathology in Chicago, October 9, 1993, and on receipt of the American Psychological Association Award for Distinguished Professional Contributions to Knowledge in Los Angeles, California, August 14, 1994.

I am grateful to Mark F. Lenzenweger and Leslie J. Yonce, to the former for suggestions to improve the final draft of this article and to the latter for preparing the figures.

Correspondence concerning this article should be addressed to Paul E. Meehl, Psychology Department, University of Minnesota, N218 Elliott Hall, 75 East River Road, Minneapolis, MN 55455-0344.

 0003-066X/95/$2.00
Vol. 50, No. 4, 266-275

group therapy), there are (*mn*) input–output relations that easily amount to several hundred. With a diagnostic construct as a mediator, we need only investigate *m* diagnostic validities and *n* prognostic or therapeutic indications, yielding a total of (*mn*) empirical correlations to be researched. The savings in research studies can approach an order of magnitude (cf. Meehl, 1959). Basic research in psychopathology will proceed differently, given a corroborated taxonic conjecture (e.g., a search for a specific gene's biochemical effect vs. clarifying a polygenic etiology).

The usefulness of categories in both research and clinical application impels us to seek high reliability; otherwise the relations are not generalizable across clinical settings. But this can produce an obsession with reliability instead of construct validity, and it can foster a belief in operational definitions of entities that are not literally operational when their explication and use are scrutinized (Faust & Miner, 1986; Meehl, 1986a). A set of disjunctions and conjunctions (as in the *Diagnostic and Statistical Manual of Mental Disorders* [*DSM*], "two or more of the following" or "at least one of the following") should be based on objective evidence of the construct validity of the various possible patterns, which for just a dozen symptoms is over 4,000. Truth by committee is initially unavoidable, but we should not persist in that. Revising operational criteria on the basis of committee discussion of clinical experience or statistical research lacking powerful taxometrics is not good science. Revisions based on such things as clinical impressions, the persuasiveness of arguments, the profession, prestige, fluency, and social dominance of committee members will sometimes improve the criteria, but will sometimes make them worse. A knowledge claim should bring credentials that it is genuine knowledge. Given 10 people guessing the distance to the moon, one might say 10,000 miles, another might say 1,000,000 miles, and a lucky one might get it right at 238,000 miles. This correct guess does us no good, either as theorists or practitioners, unless we know what credentials that guess brings compared with others.

Psychologists criticizing the *DSM* sometimes conflate two issues. The first is whether the categorical model is a good one and for which alleged syndromes or entities. It may be good for some and not for others. Second, although the categorical model may be appropriate for some mental disorders, is it being properly implemented? The threshold question, "Are there any real taxa in psychopathology?" may be answered affirmatively but the sequential question, "Are we identifying them properly by present methods?" may be answered negatively. The catch phrase *medical model* has no utility for increasing theoretical comprehension and very little utility for improving clinical practice.

Most critics don't even get it right. On the one hand, they fault the *DSM* committees for using the medical model, which is dogmatically assumed to be inapplicable. But they ignore the fact that entities in the advanced specialties of medicine are not constructed like the *DSM* categories. The advanced-science medical model does not identify disease taxa with the operationally defined syndrome; the syndrome is taken as evidentiary, not as definitory. The explicit definition of a disease entity in nonpsychiatric medicine is a conjunction of pathology and etiology and therefore applies to patients who are asymptomatic (which is why, e.g., one can have a silent brain tumor or a staghorn kidney that never causes trouble during life and is only found postmortem). Perhaps we cannot blame psychologists ignorant of medicine for making this mistake, when some psychiatrists who are passionate defenders of the *DSM* don't understand how far it deviates from the optimal medical model. Accepting operationism (an erroneous philosophy of science) and the pseudomedical model (definition by syndrome only) engenders a wrongheaded research approach, unlikely to pay off in the long run.

Discouragement with the debatable revisions by committee leads many to conclude that scientific categories are purely "conventional," with the unfortunate connotation that scientific categories are whimsical, arbitrary, and not subject to rational argument and empirical evidence. Fictionism about theoretical constructs is a fallacious inference from the obvious fact that human beings write definitions and invent theories. I blame the logical positivists a bit for this because their initial emphasis on definition was a stipulation as to the use of words. But they were not as naive about it as psychologists who rely on this truism to draw a false conclusion. Gustav Bergmann, Kenneth Spence's in-house philosopher, used a simple example to refute this notion of totally conventional arbitrariness. He spoke of the Bergmann Index, operationally defined as IQ squared, divided by the cube root of body weight, and he pointed out that no one would offer such a definition for scientific purposes. The mixup here is between "humans write definitions" and "humans write them arbitrarily" (i.e., without any idea about the way the world works).

The most important and powerful kind of definition in theoretical science is not operational but contextual or implicit, the meaning of a theoretical concept provided by its role in the postulated law network. That is why it is possible for a set of theoretical statements to both define and assert, contrary to what some critics of Cronbach and Meehl's (1955) article on construct validity have alleged.

Biological taxa are defined with words that biologists choose, relying on the relevant morphological, physiological, ecological, and ethological facts. We admire Linnaeus, the creator of modern taxonomy, for discerning the remarkable truth—a "deep structure" fact, as Chomsky might say—that the bat doesn't sort with the chickadee and the whale doesn't sort with the pickerel, but both are properly sorted with the grizzly bear; whereas Pliny the Elder had it the other way around. We do not say we have merely chosen the conventional definition of an 18th-century Swede in preference to that of a 1st-century Roman.

It must be obvious that I am not a scientific fictionist but a scientific realist. I see classification as an enterprise

that aims to carve nature at its joints (Plato), identifying categories of entities that are in some sense (not metaphysical "essentialist") nonarbitrary, not man-made. The verbal definition of them once we have scientific insight is, of course, man-made, a truism that does not prove anything about ontology or epistemology. There are gophers, there are chipmunks, but there are no gophmunks. Those two species would be there whether any human being had noticed them or christened them (Meehl, 1992a; Meehl & Golden, 1982).

Biases Against Latent Entities

Associated with the pseudomedical model and out-of-date pseudo-operationism is a fear of inferred theoretical entities—latent entities, I shall call them. They are not, despite Tolman's (1932) claim, immanent in the data, but they are inferrable from the data if one does it right. Some of them are intrinsically unobservable at the molar behavior level, although perhaps observable by sciences lower in Comte's pyramid (e.g., neurochemistry). Others, such as the positron, are unobservable in principle. An important kind of latency is unobservability not in principle but only in fact, such as a macro-object historical event which no theorist was present at the time to observe. That kind of latency applies in psychopathology when we try, as in psychoanalysis, to reconstruct a life-historical event from the verbal and gestural behavior of the patient on the couch.

Some superoperational psychologists talk as though inferring theoretical entities were somehow methodologically sinful. But several respected subfields take for granted the latent–manifest distinction. One cannot do theoretical genetics without distinguishing dominant and recessive genes, degrees of penetrance, epistatic effects, and pleiotropic markers—all of which concepts presuppose that a gene can be present but its phenotypic indicator absent. Classical psychometrics involves factors, true scores, latent variables, threshold values in multidimensional scaling, or classical item discrimination theory—all concepts not explicitly defined by the items. Rat experiments on the latent learning controversy would be quite meaningless if one could not legitimately invoke something internal that the rat had acquired but was not currently manifesting in its choice behavior. All cases of silent disease in organic medicine are like this. Finally, some of us accept parts of Freud's theoretical edifice; however, unconscious processes, impulses masked by one of the 20 defense mechanisms, and the whole psychoanalytic procedure of discerning hidden guiding themes in the patient's associations are absurd if every theoretical entity must be operationally defined.

The taxometric procedures I have invented make no theoretical sense outside of my realist philosophy of science (Meehl, 1990a, 1990b, 1993a, 1993b; cf. Feigl, 1950; Hacking, 1983; Leplin, 1984; Newton-Smith, 1981; Phillips, 1992; Popper, 1983; Salmon, 1984; Watkins, 1984). I suppose a fictionist could find them useful, but a consistent conventionalist would be incapable of understanding them, as the procedures ask an empirical question rather than invite some arbitrary stipulation about the use of words.

Confusion About Sharpness of Group Boundaries

Although the basic taxonomic question ("Is the pattern of observed relationships corroborative of a latent taxon or of latent dimensions or a mix of the two?") is a factual rather than a semantic matter, there are some unhealthy semantic habits that make this factual question harder to answer than need be. An example is the widespread careless definition of a category or class concept as involving sharp distinctions or clear-cut boundaries. Neither the mathematics nor numerous examples in the life sciences (where the causality is quite well understood, perhaps by experimental rather than taxometric methods) show that a real taxon always entails quantitative sharpness, such as a step function in one of the taxon indicators. Empirically, this is rarely the case, even in biological and medical examples. Psychologists who think there must always be clear-cut boundaries are mixing the indicators with the latent taxon they indicate. The distinction between qualitative and quantitative, or between a quantitative variable having a step function and one that behaves smoothly (even in the discriminating region of interest), can occur in all four combinations. One can have a sharp latent taxon, defined by a specific dichotomous (present or absent) causal factor (e.g., the Huntington mutation) that is indicated by quantitative variables. The specific etiological agent is a yes or no matter—one either has that mutation at the Huntington locus or not—but the individual differences in clinical features, such as age of onset, are polygenically determined, as shown by the high correlation of age of onset between siblings who have both received the dominant gene (completely penetrant, if one survives the morbidity risk period). Is the Huntington syndrome "sharp?" Of course not. A few patients develop symptoms sufficiently late in life, so that when a member of a pedigree dies fairly young, we do not know whether that person carried the gene or not.

Organic diseases with a clear-cut specific etiology, such as a specific pathogenic microorganism, often give rise to fever as a symptom, but a patient's temperature is a quantitative indicator variable. A Minnesota Multiphasic Personality Inventory (MMPI) item scored for social introversion is a dichotomous fallible indicator of the individual's position on a latent dimension. A pathognomonic sign (positive Wassermann) of an organic disease entity (syphilis) is a dichotomous indicator of a latent category. A psychometric test score loaded with a factor is a quantitative indicator of a quantitative latent variable (e.g., Wechsler Adult Intelligence Scale [WAIS] subtest score loaded with *g*). These examples illustrate the disutility of terms like *sharp* in the metatheory of taxa and dimensions. Both the latent entity and its manifest indicators can be either qualitative or quantitative, and all degrees of overlap between quantitative indicators' distributions can occur, so that the usual talk about *sharp edges* results in conceptual and empirical muddle.

A weaker form of this confusion about sharpness is the nearly ubiquitous claim that a quantitative indicator of a latent taxon must be bimodally distributed. Indicator bimodality is neither a necessary nor sufficient condition for latent taxonicity, as has long been known (see, e.g., Murphy, 1964). Two latent distributions of equal variance and a mean difference of two standard deviations will just barely yield bimodality when the base rate is one half. For a fixed mean difference, reduction in the base rate shifts the composite curve from platykurtosis to leptokurtosis with a correlated rise in skewness, a complicated exchange in the manifest distribution that remains to be thoroughly investigated. Despite Murphy's findings, I think that bimodality is strongly suggestive of taxonicity and that either marked platykurtosis or skewness is somewhat indicative, but none of these can be considered criterial.

Those who focus on dichotomous specific etiology as the most interesting kind of taxonicity must keep in mind that specific etiology is a strong, special kind of causality located far out on a metadimension of causal influences differing in specificity and strength (Meehl, 1972, 1977). It is not the only source of taxonicity. Statistical taxa can be generated by a step function on a quantitative variable (e.g., vitamin deficiency, a polygenic system influencing *g*) or on a composite of such. If the social environment (e.g., religious sect, training program, family, political regime) imposes a correlated set of deviations on several quantitative causal factors, a strong outcome taxon may result.

Coherent Cut Kinetics

Taxometrics may be defined as that branch of applied mathematics that deals with the classification of entities (Meehl & Golden, 1982). It does not matter whether the entities are mental patients, skilled tradesmen, species of honeybees, or kinds of rocks. The Classification Society includes psychologists, geologists, and even astronomers. Taxometrics is easy when one has a gold standard criterion, such as the pathologist's report in organic disease, or a pragmatic measure, such as how much insurance an agent sells per year. Linear discriminant function or other more complicated criterion-based statistics are then appropriate. Because there is no gold standard criterion in psychopathology at present, even for disorders known to be genetically influenced, psychologists require the difficult kind of taxometrics that I call *bootstrap taxometrics.* (Cronbach and Meehl first used the word *bootstrap* in this context in 1955, before its proliferated use among the statisticians and philosophers.) Lacking a gold standard criterion, the only rational basis for inferring the existence of a taxonic entity, a real class, a nonarbitrary natural kind, must lie within the pattern displayed by the presumed indicators of the conjectured taxon. In the field of psychology, as in all of the life sciences, these indicators are almost always fallible. The most widely known approach to bootstrap taxometrics is the cluster algorithms, the classic treatise being that of Sneath and Sokal (1973). Cluster algorithms have, by and large, not lived up to expectation in the social sciences. I have elsewhere (Meehl, 1979) listed eight plausible explanations for why they have not turned out to be powerful and will not discuss that further here. For many years, I have been developing new taxometric procedures for analyzing the genetics of schizophrenia, but also for a broader application. My coherent cut kinetics method covers several mathematically related procedures, and I shall briefly describe two of them. The mathematics speaks for itself and the Monte Carlo runs are encouraging, but ultimately the test of a taxometric method is its ability to solve real research problems, and readers are encouraged to try these procedures on their research problems.

The essence of any scientific procedure is classifying and quantifying in such a way as to reveal order in the data. This optimizing of orderliness stands out more strikingly when we are doing bootstrap taxometrics because of the absence of a gold standard criterion. But a close look at any of the more developed sciences, especially in their early stages, shows that they also are engaged in a bootstraps operation, whether they describe it that way or not. In psychology, such diverse thinkers as Allport, Cattell, Freud, Murray, Skinner, and Thurstone—who one sometimes thinks could hardly have had a meaningful conversation with each other because of their vast differences in method and substance—all had the maximizing of orderliness in the material as their guiding principle, and all wrote explicit methodological passages to that effect (Meehl, 1986b). Each of the different statistical procedures in my overall method is motivated by that basic scientific principle.

Figure 1
Differences in the ϕ-Coefficient When Continuous Criteria Are Cut at Different Levels

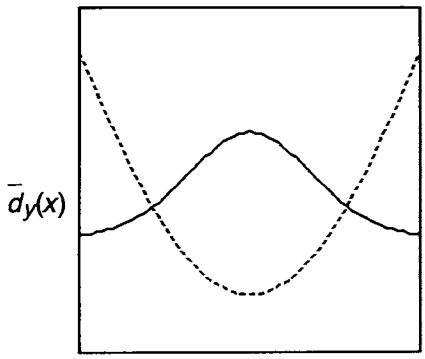
		Neck pain: extreme	Neck pain: moderate or none	
Temperature	≥105°	18	2	20
	<105°	3	77	80
		21	79	100

ϕ=.85

		Neck pain: moderate to extreme	Neck pain: none	
Temperature	≥100°	25	20	45
	<100°	15	40	55
		40	60	100

ϕ=.29

Figure 2
MAMBAC Error-Free Curve Shape for P = .50, 2 σ Separation on Each Variable, and No Nuisance Covariance



Note. Solid line is the taxonic situation; dashed line is the nontaxonic situation when r_{ij} = .50. Reprinted from "Taxometric Analysis: I. Detecting Taxonicity With Two Quantitative Indicators Using Means Above and Below a Sliding Cut (MAMBAC Procedure")*,* by P. E. Meehl and L. J. Yonce, 1994, Psychological Reports, 74, p. 1073. Copyright 1994 by Dr. C. H. Ammons and Dr. R. B. Ammons (Editors and Publishers).

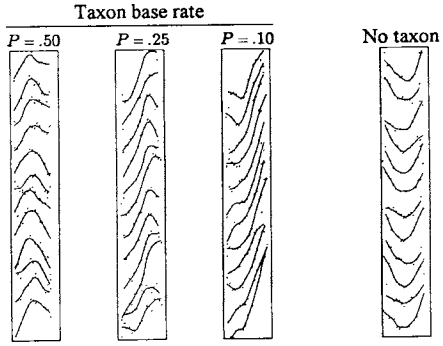
MAMBAC

Let me begin with the simplest, MAMBAC (acronym from the phrase "mean above minus below a cut"; see Meehl & Yonce, 1994). Consider a simple example from organic medicine, meningitis. Extreme pain upon anteroflection of the neck is found in meningitis, along with a markedly elevated temperature. Imagine a clinical population containing a sizable subset of patients with meningitis, another subset with a mixture of other various organic diseases, plus some individuals without organic disease. We could define a dichotomous sign, high fever, as, say, over 105°, and a dichotomous symptom—a painful, stiff neck. If we set up a fourfold table (see Figure 1) showing the concordance between these two dichotomous indicators, the patients with meningitis will occupy the (++) concordant cell and almost everybody else will be in the (−−) concordant cell, yielding a ϕ-coefficient close to 1. If we choose a much lower value for the fever cutting score (say, a temperature of 100°) and require only a mildly stiff neck, what will happen to the table? Patients with various other diseases will have temperatures that high, and the slight neck stiffness will be due to a miscellany of conditions including not only meningitis but also cervical arthritis, influenza, a sleeping posture, chill, or whatever. Because this melange of other conditions will often have one of the dichotomous signs without the other, we will find numerous tallies in the discordant cells and the ϕ-coefficient will be markedly lowered. That intuition motivated the MAMBAC statistic.

Consider a quantitative indicator variable y that has sizable validity for discriminating a taxon and a complement class. If we partition a mixed population into these two categories, the difference between their y means will be as large as it can be made by any partition that does not rely on the y values themselves; the mean difference on that optimal partition will be the difference between the means of the taxon and complement classes. Suppose we begin to interchange cases, missorting some members of the taxon into the complement group and conversely, mixing some members of the complement into the taxon group, again, in ignorance of the y values. The effect of that intermixing procedure will be to reduce the observed difference $\overline{\Delta y}$.

Suppose x and y are independent within the categories so that their correlation is attributable only to the taxonic structure. If we now sort subjects on the basis of x, there should be a relationship between how well x partitions the group and the observed difference between the groups on y. So we define a statistical function of the x cut as $\bar{d}_y(x)$, which is the difference between the mean y of the cases above the x cut and the mean y below that cut. Examining the behavior of that $\bar{d}_y(x)$ statistic as the cut moves along the x indicator, we find it has a maximum (see Figure 2). This hump will be at the x cut that gives the best partitioning of the group on y. The important point is that the convex upward appearance of a graph of $\bar{d}_y(x)$ (a hump) indicates a latent taxonicity. If the latent structure is nontaxonic (i.e., the observed correlation between indicators x and y is produced by a latent quantitative factor rather than by a pair of partially overlapping categories), it can be shown that the graph is concave upward, resembling a dish or saucer rather than a hump. The graph shape answers the threshold question that we must resolve before we even discuss how to specify a taxon, namely, that there is a taxon rather than a dimensional factor. If we have four continuous variables, we have six pairs to study, but each can be worked in either direction (i. e., x as input, y as output, then the converse) for a total of 12 graphs. Figure 3 shows sets of graphs generated with artificial (Monte Carlo) data for taxonic

Figure 3
MAMBAC Monte Carlo Curves for Different Base Rates and for the Nontaxonic Situation



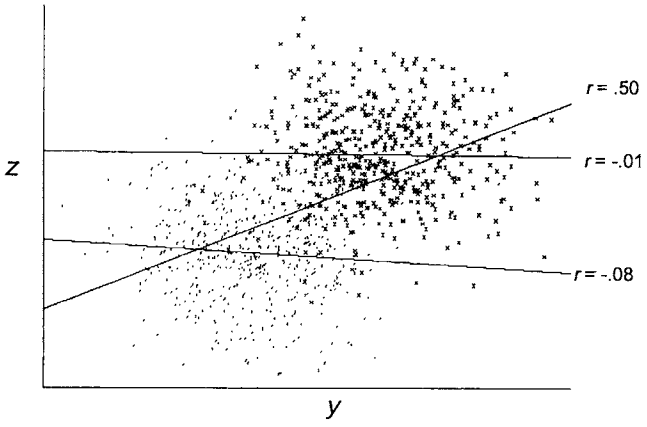
cases with different base rates and for the factorial case. Notice that the high part of the taxonic MAMBAC curve shifts to the right with lower base rates, becoming a right-end cusp when P = .10. Panels such as these can be sorted with very high accuracy (98% correct or better), even by persons with no social science background (Meehl & Yonce, 1994).

Having answered the threshold question of taxonicity, the next question is, "What is the base rate?" The MAMBAC procedure allows us to estimate the base rate, to infer the taxonic separation, and to estimate the values of the latent taxon and complement means (Meehl & Yonce, 1994). One can then assign individuals to taxon or complement membership with Bayes's theorem. MAMBAC has been used by Waller, Putnam, and Carson (1994) to detect a taxon of pathological dissociation.

MAXCOV

The MAXCOV–HITMAX procedure (Meehl, 1973) relies on a different aspect of orderliness and was devised for three indicator variables (but cf. Grove & Meehl, 1993, for a recent two-indicator variant). Consider the taxon of biological gender. Females will have a higher mean score on a good test of Murray's nurturance need and also on the Minnesota Clerical Test. Assume, as seems plausible, that there is no appreciable correlation between clerical ability and nurturance within gender groups. Then the covariance within each category will be zero and the regression line of y on z or z on y will be flat (see Figure 4). But if we mix the two groups, the systematic difference between the means will lead to a significant correlation and a nonflat regression line (although a better fit will be nonlinear, with a jog somewhere in the middle region). The basic algebra is shown in the general covariance mixture theorem,

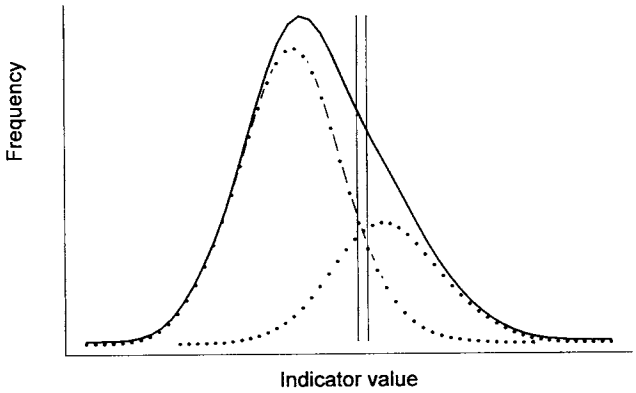
Figure 4



Note. Scatterplot showing negligible nuisance covariance (i. e., correlation within either the complement or the taxon group) but sizable covariance in the total sample as a result of separation between the complement and taxon groups.

Figure 5

A Smoothed Frequency Distribution (Solid Line) and the Latent Frequency Distributions for the Complement and Taxon Groups That Comprise the Total Sample



Note. Vertical lines mark the hitmax interval. (These curves were drawn from a taxonic Monte Carlo sample of 1,000 with a base rate P = .30.)

$$cov_{yz}(x) = p\ cov_t + q\ cov_c + pq\ (\bar{y}_t - \bar{y}_c)(\bar{z}_t - \bar{z}_c).$$

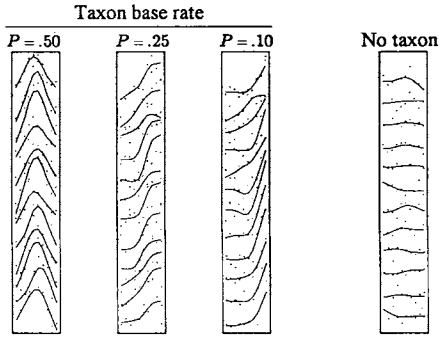
If there is negligible nuisance covariance within the categories, the observed covariance of y and z is a function of the amount of taxon mixture (i.e., pq) multiplied by the product of the mean separations $\Delta\bar{y}\Delta\bar{z}$,

$$cov_{yz}(x) \simeq pq\ (\bar{y}_t - \bar{y}_c)(\bar{z}_t - \bar{z}_c).$$

If we have a third indicator x that has validity, we can arrange subsamples along the x dimension, and the proportion of taxon members in each x interval is a monotone increasing function of x. The product pq is a maximum for an even mix, which occurs in the interval surrounding the x cut that maximizes the correct classifications (see Figure 5). I call this the *hitmax interval*, where $p = q = 1/2$. The MAXCOV graph will show a

Figure 6

MAXCOV Monte Carlo Curves for Different Base Rates and for the Nontaxonic Situation



clear peak at this location. Having located the hitmax interval by finding the peak of the *yz* covariance graph, we know that in that interval the product $pq = 1/4$. Working from that, we can estimate the taxon frequencies in the other intervals and the overall base rate (see Meehl, 1973; Meehl & Golden, 1982).

If we have three variables to work with, we can choose each in turn as the input and study the covariance of the other two. With four variables available, we can pick a triad in four ways; for each triad, there are three ways to select the input variable, producing a total of 12 MAXCOV graphs and 12 estimates of the base rate. The Monte Carlo examples in Figure 6 show very clear shape differences for taxonic and nontaxonic situations, although in this procedure the factorial situation does not give a dish (as it does with the MAMBAC procedure) but a flat graph. I sorted several hundred such panels into taxonic and nontaxonic categories with over 95% correct decisions, and a nonpsychologist did as well.

If the base rate is near .50, the peak of the MAXCOV graph will be centrally located. In situations with base rates lower than .50, the peak moves to the right, the extreme case occurring when the base rate is so small that there is no local maximum in a mathematician's sense (because the curve has no room to bend downward again), and instead of a peak there is a cusp at the extreme right. The MAXCOV procedure (or variants of it) has been used in at least 17 studies (see various applications cited in Meehl, 1992a, p. 135; Harris, Rice, & Quinsey, 1994; Korfine & Lenzenweger, 1995; Lenzenweger & Korfine, 1992, in press; Nicholson & Neufeld, 1994; Waller et al., 1994).

Preferably, MAMBAC and MAXCOV employ quantitative indicators of wide range—which psychometrically requires many items—for both input and output, and researchers are strongly urged to select or construct such continuous measures. However, if there is at least one quantitative indicator for input, it is possible to employ dichotomous signs as output, since algebraically a difference between two proportions ($p_A - p_B$) is a mean difference, and the numerator ($p_{ij} - p_i p_j$) of a ϕ-coefficient is a covariance. The limitations of using dichotomous output indicators remain to be investigated.

Consistency Tests

Consistency tests may be roughly defined as comparisons of numerical values inferred from the procedures to see whether they cohere. Comparing two or more estimates of the same latent value (e.g., base rate) reached by different epistemic paths yields three kinds of consistency tests: First, we can use the same statistical procedure but different indicators. Thus, with multiple indicators (x, y, z, and v), we can look at the MAMBAC graphs from (x,y) and the graphs from (z,v)—both pairs run bidirectionally—to see whether they are all taxonic; in addition, we can ask whether the base rates estimated from (x,y) and (z,v) agree within tolerance. Second, with the same set of indicators but different procedures, we can study indicators (x,y), (x,z), and (y,z) with MAMBAC, and we can study the triad (x,y,z) with MAXCOV. Third, a severe test of the taxonic conjecture is comparing the results of different indicator sets through different procedures, running MAMBAC on indicators (x,y) versus MAXCOV on indicators (z,u,v). Coherence among such nonredundant estimates of base rate and latent means provides strong corroboration.

A more complex kind of consistency test involves theorems derivable from the postulated taxonic model, showing various mathematical relations between sets of indicator values and various latent values. I give only one example: In the MAXCOV procedure, the covariance mixture theorem, which we apply in successive class intervals along the input variable, also holds for any set of cases and therefore, in particular, for the whole sample. We can write three grand covariances as functions of the base rate and the latent validities (mean separations):

$$\mathrm{cov}_{xy} = PQ\,\overline{\Delta x}\,\overline{\Delta y},$$

$$\mathrm{cov}_{xz} = PQ\,\overline{\Delta x}\,\overline{\Delta z},$$

and

$$\mathrm{cov}_{yz} = PQ\,\overline{\Delta y}\,\overline{\Delta z}.$$

For a taxonic situation, these should hold, within tolerance (from Meehl & Golden, 1982, p. 165, Equation 24).

It is a mistake to think of consistency tests in bootstrap taxometrics as a sort of luxury, pleasantly reassuring if one is so lucky as to get it. Consistency tests are an absolute necessity in bootstrap taxometrics, and the more available the better. My emphasis on coherency springs from my neo-Popperian philosophy of science, the need for having risky tests (O'Hear, 1980; Popper, 1959, 1962, 1972; Schilpp, 1974). If the reader does not care for Popper and is, like most psychologists, more of an inductivist, the same conclusion flows from philosopher Wesley Salmon's analysis of "damned strange coincidence" (Nye, 1972; Salmon, 1984, and personal communication, June, 1980). The history of science makes it clear that risky Popperian tests or Salmonian strange coincidences play a major role—perhaps the biggest single role—in corroborating scientific theories. To get risky tests or Salmonian coincidences, a theoretical conjecture must predict, not merely fit a theoretical model to the data after the fact. A strong scientific theory allows prediction of the data, sometimes even of a particular numerical value. Weaker theories, such as in psychopathology, do not permit this kind of prediction, but they do predict that certain relationships should obtain within the data; that is, from one set of numbers found in the data, it should be possible to predict another set of numbers, within tolerance. The absence of consistency tests in psychometric procedures leads to excessive reliance on judgment and thereby to a lack of consensus among critical scientists.

Idealizations and Robustness

All science idealizes theoretical constructs, which in turn leads to an idealization in the formalism. How is our formalism false if taken literally? First, the theory is in-

complete. Therefore, derivations in the formalism requiring what philosophers call the ceteris paribus ("other things being equal") clause do not strictly go through, because the ceteris paribus clause is never literally true. Second, there are simplifying auxiliary conjectures, such as the assumptions of normality, linearity, and homoscedasticity that are rarely literally true. The coherent cut kinetics method uses the convenient auxiliary conjecture of independence within the categories, which we hope is close to true but we know is not literally true, because in the life sciences everything is correlated with everything (cf. Meehl, 1978, 1990a, 1990e). Finally, the mathematics used is continuous, whereas the data are always discontinuous, no matter how finely grained, and there is no reason to believe that the underlying dimensions are continuous. Thus, for example, when we find maxima and minima by zeroing a derivative we are idealizing this physical state of affairs. The idealization of the theoretical constructs and the associated idealization in the formalism entails approximateness of numerical values, which, of course, we already start with because observational measurements are subject to error.

We idealize partly because we don't know about everything and therefore want to begin by testing a weaker theory that we know is incomplete. We also idealize to permit a tractable formalism. This raises the problem of robustness. It is not fruitful to ask, "Do the data depart from the theory?" because we know they do. They deviate in the literal sense because of sampling error (and only this can be examined by conventional significance testing) and because auxiliary conjectures are not exactly true. Sometimes our initial conjecture about the underlying structure is incorrect. For instance, there may be no taxon, and the observed correlations are generated by an underlying, quantitative factor having a single unimodal latent distribution (in which case my method will refute the initial conjecture).

The most important idealization in my method is that of negligible nuisance covariance within the latent categories, which is represented in the math by assigning it a value of zero. The limits on robustness are still under investigation, but preliminary results suggest that correlations not exceeding .25 or .30 will do very little damage, and larger values are tolerable if they are approximately equal in the two categories. Psychologists should not be fretful about this idealization, considering that in the "soft fields" like psychopathology and personality theory we usually work hard to get correlations up to .40 or .50! Procedures for explicitly taking the nuisance covariances into account are under investigation (Meehl, 1995).

Why use Monte Carlo methods instead of seeking rigorous analytical solutions to the problem of tolerance and robustness? The short answer is this: We idealize in the first place because the mathematics for the literally correct state of affairs, including our conjectures regarding it in all details, is intractable. That is the common experience, even in fields such as physics and chemistry. Thus, physicists almost always work with linear differential equations, even when they have reason to think that the complete picture would be otherwise. If one tried to answer the question of numerical tolerance and robustness analytically, it would involve deriving expressions in the formalism for the discrepancy between the idealized and the literally true mathematical statements. But, if the literally true formalism is intractable, then any mathematical expression of the difference between the two, which will necessarily be comparing the idealized expressions with the nonidealized expressions, will be a fortiori intractable.

Psychologists trained, as I was, in conventional Fisherian statistics may ask, "Why do we want several estimates of the same parameter? Don't we just want to get a maximum likelihood estimator?" No. Space does not permit a detailed discussion of this, but the essential point is that in agronomy, the paradigm case of Fisherian statistics, the variables are all observational variables such as bushels of wheat or pounds of fertilizer applied, and the source of error is random sampling fluctuation in the seed and the soil. There are no inferred theoretical entities such as positrons, libido, habit strength, or major genes. The most important source of error for sciences using theoretical entities is not the random sampling fluctuation of observational measures (which can always, in principle, be taken care of by replication and by increasing sample size); rather, it is the transition between an accepted statistical value inferred from the data and the surplus meaning involved in assertions concerning the theoretical entities (see MacCorquodale & Meehl, 1948; Meehl, 1978, 1990a, 1990e).

The classical paradigm case of multiple epistemic paths in the history of physics is the determination of Avogadro's number, the number of molecules in one gram molecular weight of a substance. By World War I it was already inferable from some 13 different databases, qualitatively distinct and nonoverlapping, ranging from the fact that the sky is blue to the statistical distribution of displacements of a Brownian particle. The derivation chains from statements about the molecules to these various theorems about observational data diverge almost immediately, both in the interpretive text and in the formalism. If there are not any molecules, it is incredible that these 13 methods of counting them should give roughly the same result, namely, around 6×10^{23} (Nye, 1972; Salmon, 1984; but cf. Carrier, 1991, and Meehl, 1992b). This convergence led the skeptic Poincaré to abandon his fictionist view and to state that if there are 13 nonredundant ways of counting something, and they give the same answer, then there must be something that is being counted. It is inappropriate to ask which of these 13 approaches is the maximum likelihood estimator of the number of molecules. The Fisherian inference model simply does not fit this situation. But it would be equally foolish to have done an F test (had it existed in those days) on the 13 values, because that would have shown statistically significant differences and led, quite wrongly, to the conclusion that molecular theory was false.

Importance of Valid Indicators and Large Samples

I take a strong stand against using measures of weak validity. We will explore how large samples must be before

extremely weak separations can be used in coherent cut kinetics procedures, but for now, I think researchers should not use indicators with a mean separation of less than 1.25 standard deviations. This is not unduly optimistic. As a rule of thumb, I suggest that one probably cannot do good taxometric research with indicators that are poorer than the weakest MMPI scales as validated against old-fashioned pre-*DSM–III* diagnoses (*Diagnostic and Statistical Manual of Mental Disorders, 3rd ed., American Psychiatric Association, 1980*). For a base rate of .50, I favor requiring 75% correct classifications. No MMPI scale is worse than that.

Taxometric research also requires larger samples than psychologists are accustomed to using. I recommend a sample size of 300 or more. (Although both real and Monte Carlo data sometimes show good performance for sample sizes as small as 100, I do not approve of using such small samples for taxometric research.) A researcher who does not have a sufficient number of participants and valid measuring instruments to do taxometric research should do something else. Physicists, chemists, astronomers, geneticists, and epidemiologists have long recognized that some questions can only be answered with large samples and precise instruments, and it is time for psychologists to adopt the same attitude.

Role of Theory in Taxometrics

Taxonicity is defined and identified by statistical patterns of indicators revealing a latent formal structure, leaving open the substantive interpretation. No statistical procedure is self-interpreting as to the nature of the inferred theoretical entities and their causal relations. Equating *taxon* with *disease entity* with *specific etiology* with *germ or gene* is unwarranted and further intensifies psychologists' antitypological prejudice. Many taxa are environmental mold types, such as religious and political ideologies. The political taxon *Trotskyist* is a more tightly knit syndrome than any in the *DSM* or many organic diseases in internal medicine. Other than biological species, the largest number of taxa is found in the *Dictionary of Occupational Titles* (Department of Labor, 1977), in which there are over 20,000 entries.

My neo-Popperian philosophy of science engenders a preference for theory-motivated selection of candidate indicators. But one should not try to impose one's metatheory on others, so although I advocate testing taxonic conjectures one by one (is schizotypy a taxon? cyclothymia? Cleckley-Lykken psychopathy? hysteria?), I also wish to assist researchers of inductive bent who confront a large batch of k miscellaneous variables—signs, symptoms, personality traits, life history facts, interview ratings or checklists, psychometric and psychophysiological measures—and prefer to analyze them through blind inductive scanning of their statistical relations. We have not as yet written a program for doing this with the coherent cut kinetics method, but a researcher can easily do it by sequencing MAMBAC and MAXCOV. First, compute the conventional Pearson pairwise correlations between all $\binom{k}{2}$ pairs of variables and identify the related pairs. These pairs are taxonic indicator candidates, but may instead reflect factor loadings on one or more (noncategorical) latent dimensions. MAMBAC is used to identify the taxonically generated pairs. Finally, triads from overlapping sets of taxonic pairs are analyzed by MAXCOV. The numerous consistency tests available in these multiple analyses provide strong corroboration of the multiple taxonic conjecture. The correlated pairs shown to have a nontaxonic latent source can of course be subjected to an appropriate dimensional procedure such as factor analysis. We proceed in this taxon-then-factor sequence because factor analysis cannot answer the threshold taxonic question. To think that it can is a simple mathematical mistake, inasmuch as a batch of items or scales strongly discriminating a taxon will always appear as a factor.

I previously resisted adapting my taxometrics to all such atheoretical inductive scannings because of the relatively poor performance of cluster analysis in psychology, and also because the majority of biologists have not accepted numerical taxonomy as the way to solve their classification problem either. But I now think I was mistaken in this reasoning, and that the main weakness of the favored cluster methods has a different source. The first step in those methods is to calculate a matrix of interindividual distances (or similarities), thus, "How close is honeybee i to honeybee j in the descriptor space?" Mathematical statisticians complained early on that nondifferentiating variables contributing to this pairwise distance measure would swamp the minority of differentiating variables, a rigorous form of the more intuitive complaints of traditional taxonomists (such as Ernst Mayr) that not all variable features should be considered equally important. I do not believe that this objection has been satisfactorily answered. But because my method does not at any stage rely on such an interindividual distance matrix, the objection does not apply. The pairs, triads, and larger sets of indicators revealed by coherence of the procedures are not affected by the copresence of other variables that either are uncorrelated or are correlated because of nontaxonic factors.

Conclusion

From the empirical studies to date and extensive Monte Carlo runs, I conjecture that I have solved the taxometric problem. If a researcher is correct in conjecturing the existence of a latent taxon and has halfway decent indicators, the coherent cut kinetics method will show that it is taxonic, estimate the base rate accurately, locate the optimal cuts on indicators, estimate the validities those cuts achieve, reassure as to the model by multiple consistency tests, and provide a classification of individuals as accurate as the indicators permit. I hope this article will induce readers to put it to the test in diverse substantive research domains.

REFERENCES

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Carrier, M. (1991). What is wrong with the Miracle Argument? *Studies in History and Philosophy of Science, 22,* 23–36.

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

Department of Labor, Employment and Training Administration, Employment Service. (1977). *Dictionary of occupational titles* (4th ed.). Washington, DC: U.S. Government Printing Office.

Faust, D., & Miner, R. A. (1986). The empiricist and his new clothes: DSM-III in perspective. *American Journal of Psychiatry, 143,* 962–967.

Feigl, H. (1950). Existential hypotheses: Realistic versus phenomenalistic interpretations. *Philosophy of Science, 17,* 35–62.

Grove, W. M. (1991). When is a diagnosis worth making? A comparison of two statistical prediction strategies. *Psychological Reports, 68,* 3–17.

Grove, W. M., & Meehl, P. E. (1993). Simple regression-based procedures for taxometric investigations. *Psychological Reports, 73,* 707–737.

Hacking, I. (1983). *Representing and intervening.* New York: Cambridge University Press.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1994). Psychopathy as a taxon: Evidence that psychopaths are a discrete class. *Journal of Counsuling and Clinical Psychology, 62,* 387–397.

Korfine, L., & Lenzenweger, M. F. (1995). The taxonicity of schizotypy: A replication. *Journal of Abnormal Psychology, 104,* 26–31.

Lenzenweger, M. F., & Korfine, L. (1992). Confirming the latent structure and base rate of schizotypy: A taxometric analysis. *Journal of Abnormal Psychology, 101,* 567–571.

Lenzenweger, M. F., & Korfine, L. (in press). Tracking the taxon: On the latent structure and base rate of schizotypy. In A. Raine, T. Lencz, & S. Mednick (Eds.), *Schizotypal personality.* New York: Cambridge University Press.

Leplin, J. (Ed.). (1984). *Scientific realism. Berkeley:* University of California Press.

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55,* 95–107.

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

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

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

Meehl, P. E. (1973). MAXCOV-HITMAX: A taxonomic search method for loose genetic syndromes. *Psychodiagnosis: Selected papers* (pp. 200–224). Minneapolis, MN: 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. (1986a). Diagnostic taxa as open concepts: Metatheoretical and statistical questions about reliability and construct validity in the grand strategy of nosological revision. In T. Millon & G. L. Klerman (Eds.), *Contemporary directions in psychopathology* (pp. 215–231). New York: Guilford Press.

Meehl, P. E. (1986b). Trait language and behaviorese. In T. Thompson & M. D. Zeiler (Eds.), *Analysis and integration of behavioral units* (pp. 315–334). Hillsdale, NJ: Erlbaum.

Meehl, P. E. (1989). Schizotaxia revisited. *Archives of General Psychiatry, 46,* 935–944.

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry, 1,* 108–141, 173–180.

Meehl, P. E. (1990b). *Corroboration and verisimilitude: Against Lakatos' "sheer leap of faith"* (Working Paper, MCPS-90–01). Minneapolis: University of Minnesota, Center for Philosophy of Science.

Meehl, P. E. (1990c). Schizotaxia as an open concept. In A. I. Rabin, R. Zucker, R. Emmons, & S. Frank (Eds.), *Studying persons and lives* (pp. 248–303). New York: Springer.

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

Meehl, P. E. (1990e). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244. Also in R. E. Snow & D. Wiley (Eds.), *Improving Inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 13–59). Hillsdale, NJ: Erlbaum.

Meehl, P. E. (1992a). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality, 60,* 117–174.

Meehl, P. E. (1992b). The Miracle Argument for realism: An important lesson to be learned by generalizing from Carrier's counter-examples. *Studies in History and Philosophy of Science, 23,* 267–282.

Meehl, P. E. (1993a). Four queries about factor reality. *History and Philosophy of Psychology Bulletin, 5(2),* 4–5.

Meehl, P. E. (1993b). Philosophy of science: Help or hindrance? *Psychological Reports, 72,* 707–733.

Meehl, P. E. (1995). Extension of the MAXCOV-HITMAX taxometric procedure to situations of sizeable nuisance covariance. In D. Lubinski & R. V. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings.* Palo Alto, CA: Consulting Psychologists Press.

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

Meehl, P. E., & Yonce, L. J. (1994). Taxometric analysis: I. Detecting taxonicity with two quantitative indicators using means above and below a sliding cut (MAMBAC procedure). *Psychological Reports, 74,* 1059–1274.

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

Newton-Smith, W. H. (1981). *The rationality of science.* Boston: Routledge & Kegan Paul.

Nicholson, I. R., & Neufeld, R. W. J. (1994). *The problem of dissecting schizophrenia: Evidence for a dimension of disorder.* Manuscript submitted for publication.

Nye, M. J. (1972). *Molecular reality.* London: Macdonald.

O'Hear, A. (1980). *Karl Popper.* Boston: Routledge & Kegan Paul.

Phillips, D. C. (1992). *The social scientist's bestiary.* New York: Pergamon.

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books. (Original work published 1935)

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

Popper, K. R. (1972). *Objective knowledge.* Oxford, England: Clarendon.

Popper, K. R. (1983). *Postscript: Vol. 1. Realism and the aim of science.* Totowa, NJ: Rowman & Littlefield.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Schilpp, P. A. (Ed.). (1974). *The philosophy of Karl Popper* (Vols. 1 & 2). LaSalle, IL: Open Court.

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

Tolman, E. C. (1932). *Purposive behavior in animals and men.* New York: Century.

Waller, N. G., Putnam, F. W., & Carlson, E. B. (1994). *Types of dissociation and dissociative types: A taxometric analysis of dissociative experiences.* Manuscript submitted for publication.

Watkins, J. W. N. (1984). *Science and scepticism.* Princeton, NJ: Princeton University Press.