Journal of Consulting and Clinical Psychology
1978, Vol. 46, 806-834.

#113

Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology

Paul E. Meehl
University of Minnesota

Theories in "soft" areas of psychology lack the cumulative character of scientific knowledge. They tend neither to be refuted nor corroborated, but instead merely fade away as people lose interest. Even though intrinsic subject matter difficulties (20 listed) contribute to this, the excessive reliance on significance testing is partly responsible, being a poor way of doing science. Karl Popper's approach, with modifications, would be prophylactic. Since the null hypothesis is quasi-always false, tables summarizing research in terms of patterns of "significant differences" are little more than complex, causally uninterpretable outcomes of statistical power functions. Multiple paths to estimating numerical point values ("consistency tests") are better, even if approximate with rough tolerances; and lacking this, ranges, orderings, second-order differences, curve peaks and valleys, and function forms should be used. Such methods are usual in developed sciences that seldom report statistical significance. Consistency tests of a conjectural taxometric model yielded 94% success with zero false negatives.

I had supposed that the title gave an easy tipoff to my topic, but some puzzled reactions by my Minnesota colleagues show otherwise, which heartens me because it suggests that what I am about to say is not trivial and universally known. The two knights are Sir Karl Raimund Popper (1959, 1962, 1972; Schilpp, 1974) and Sir Ronald Aylmer Fisher (1956, 1966, 1967), whose respective emphases on subjecting scientific theories to grave danger of refutation (that's Sir Karl) and major reliance on tests of statistical significance (that's Sir Ronald) are, at least in current practice, not well integrated—perhaps even incompatible. If you have not been accustomed to thinking about this incoherency, and my remarks lead you to do so (whether or not you end up agreeing with me), this article will have served its scholarly function.

I consider it unnecessary to persuade you that most so-called "theories" in the soft areas of psychology (clinical, counseling, social, personality, community, and school psychology) are scientifically unimpressive and technologically worthless. Documenting that statement would of course require a considerable amount of time, but you can quickly get the flavor by having a look at Braun (1966); Fiske (1974); Gergen (1973); Hogan, DeSoto, and Solano (1977); McGuire (1973); Meehl (1960/1973a, 1959/1973f); Mischel (1977); Schlenker (1974); Smith (1973); and Wiggins (1973). These are merely some high visible and forceful samples; I make no claim to bibliographic completeness on the large theme of "What's wrong with 'soft' psychology." A beautiful hatchet job, which in my opinion should be required reading for all PhD candidates, is by the sociologist Andreski (1972). Perhaps the easiest way to convince yourself is by scanning the literature of soft psychology over the last 30 years and noticing what

This article is based on a lecture delivered at the meeting of the American Psychological Association, Washington, D.C., September 1976, on the occasion of the author's receiving Division 12, Section 3's Distinguished Scientist Award. The research reported here was assisted by Grant MH 24224 from the National Institute of Mental Health and the University of Minnesota Psychiatry Research Fund. Completion of the article was aided by a James McKeen Cattell Fund sabbatical award.

Requests for reprints should be sent to P. E. Meehl, University of Minnesota, Department of Psychology, 75 East River Road, Minneapolis MN 55455-0344.

happens to theories. Most of them suffer the fate that General MacArthur ascribed to old generals—They never die, they just slowly fade away. In the developed sciences, theories tend either to become widely accepted and built into the larger edifice of well-tested human knowledge or else they suffer destruction in the face of recalcitrant facts and are abandoned, perhaps regretfully as a "nice try." But in fields like personology and social psychology, this seems not to happen. There is a period of enthusiasm about a new theory, a period of attempted application to several fact domains, a period of disillusionment as the negative data come in, a growing bafflement about inconsistent and unreplicable empirical results, multiple resort to ad hoc excuses, and then finally people just sort of lose interest in the thing and pursue other endeavors.

Since I do not want to step on toes lest my propaganda falls on deaf ears, I dare not mention what strike me as the most egregious contemporary examples, so let us go back to the late l930s and early 1940s when I was a student. In those days we were talking about level of aspiration. You could not pick up a psychological journal—even the *Journal of Experimental Psychology*—without finding at least one and sometimes several articles on level of aspiration in schizophrenics, or in juvenile delinquents, or in Phi Beta Kappas, or whatever. It was supposed to be a great powerful theoretical construct that would explain all kinds of things about the human mind from psychopathology to politics. What happened to it? Well, I have looked into some of the recent textbooks of general psychology and have found that either they do not mention it at all—the very phrase is missing from the index—or if they do, it gets cursory treatment in a couple of sentences. There is no doubt something to the notion. We all agree (from common sense) that people differ in what they demand or expect of themselves, and that this probably has something to do, sometimes, with their performance. But it did not get integrated into the total nomological network, nor did it get clearly liquidated as a nothing concept. It did not get killed or resurrected or transformed or solidified; it just kind of dried up and blew away, and we no longer wanted to talk about it or do experimental research on it. A more recent example is the theory of "risky shift," about which Cartwright (1973) wrote, after reviewing 196 papers that appeared in the 1960s:

> As time went by . . . it gradually became clear that the cumulative impact of these findings was quite different from what had been expected by those who produced them. Instead of providing an explanation of why "groups are riskier than individuals," they in fact cast serious doubt on the validity of the proposition itself (p. 225).
>
> It is now evident that the persistent search for an explanation of "the risky shift" was misdirected and that any adequate theory will have to account for a much more complicated set of data than originally anticipated. But it is not clear how theorizing should proceed, since serious questions have been raised as to whether, or in what way, "risk" is involved in the effects to be explained (p. 226).
>
> After 10 years of research, [the] original problem remains unsolved. We still do not know how the risk-taking behavior of "real-life" groups compares with that of individuals (p. 231).

I do not think that there is any dispute about this matter among psychologists familiar with the history of the other sciences. It is simply a sad fact that in soft psychology theories rise and decline, come and go, more as a function of baffled boredom than anything else; and the enterprise shows a disturbing absence of that *cumulative* character that is so impressive in disciplines like astronomy, molecular biology, and genetics.

There are some solid substantive reasons for this that I will list here, lest you think that I am beating up on the profession, unaware of the terrible intrinsic difficulty of our subject matter. Since (in 10 minutes of superficial thought) I easily came up with 20 features that make human psychology hard to scientize, I invite you to pick your own favorites. Differences as to which difficulties are emphasized will not, I am sure, cause any disagreement about the general fact. This is not the place to develop in detail the thesis that the human mind is hard to scientize, let alone to prove it. Each of the 20 difficulties is, I am aware, debatable; and one could find competent psychologists who would either deny a difficulty's reality—at least in the form I state it—or who, although admitting it exists, would maintain that we have, or will be able to develop shortly, methods adequate to overcome or circumvent it. Each of these alleged difficulties in scientizing the human mind is sufficiently controversial to deserve a methodological article by itself. This being so, to substitute a once-over lightly (and hence inevitably dogmatic) defense of each as a real difficulty is, for those who accept it, a work of supererogation, and for the others, it is doomed to failure. I therefore confine myself to listing and explaining the problems, repeating that my purpose in so doing is to prevent the rest of my article from being taken as a kind of malicious and unsympathetic attack on psychologists (of which, after all, I *am* one!) based on an inadequate appreciation of the terrible difficulties under which we work. In a few cases I have explained at some length and replied to objections, these being cases in which a difficulty is not widely recognized in our profession or in which it is generally held to have been disposed of by a familiar (but erroneous) refutation or solution. Regrettably, some psychologists use "philosophical" arguments that are a generation or more out of date.

Since I am listing and summarizing rather than developing or proving, it seems appropriate to present the set of difficulties as follows:

1. *Response-Class Problem*

This involves the well-known difficulties of slicing up the raw behavioral flux into meaningful intervals identified by causally relevant attributes on the response side, a problem that exists already in the Skinner box (Skinner, 1938, p. 70), worsens in field study by an ethologist, and reaches almost unmanageable proportions in studying human social behavior of the kind to which clinical, social, and personology psychologists must address themselves (see, e.g., MacCorquodale & Meehl, 1954, pp. 218–231, after a quarter century still considered by some as best statement of the problem; Hinde, 1970, pp. 10–13; Meehl, 1954, pp. 40–44 and chap. 6 passim; Skinner, 1938, pp. 33–43).

2. *Situation-Taxonomy Problem*

As is well-known, the importance of an adequate classification and sampling of environments and situations has received less attention than Problem 1, above, despite emphasis by several major contributors such as Roger Barker (1968), Egon Brunswik (1955), and Saul B. Sells (1963). It seems likely that the problems of characterizing the stimulus side, even though often neglected by the profession or dealt with superficially, are about as intractable as the characterization of the response class. It is not even clear whether identification and measurement of the relevant stimulus dimensions (e.g., size) is the same task as concocting a taxonomy of "situations" and "environments," nor whether the answer to this question would quickly generate rules for an adequate statistical ecology applicable to research design. So I am perhaps lumping under this "situation-taxonomy" rubric three distinguishable but related problems. I am inclined to think that most (not all) of the current methodological controversy concerning traits versus situations is logically and mathematically reducible to this and the preceding category, since I think that traits are disposition clusters, and dispositions always involve at least implicit reference to the stimulus side; hut this is not the place to push that view.

3. *Unit of Measurement*

One sometimes hears this conflated with one or both of the preceding, but, of course, it is not the same. There are questions in rating scales and in psychometrics (as well as in certain branches of nondifferential psychology) in which disagreements persist about such fundamental matters as the necessity of a genuine interval or ratio scale for the use of certain kinds of sampling statistical inference.

4. *Individual Differences*

Perhaps the shortest way to discuss this one is to point out the oddity that what is one psychologist's subject matter is another psychologist's error term (Cronbach, 1957)! More generally, the fact is that organisms differ not only with respect to the strengths of various dispositions, but, more common and more distressing for the researcher, they differ as to *how* their dispositions are shaped and organized. As a result, the individual differences involved in "mental chemistry" are tougher to deal with than, say, the fact that different elements have different atomic numbers or that elements with the same atomic number vary in atomic weights (isotopes).

5. *Polygenic Heredity*

It is generally conceded that the measurement and causal inference problems that arise in biometrical genetics are, with some exceptions, more difficult than those found in the kind of single factor dominant or recessive gene situation on which the science of genetics was originally founded. Except for Mendelizing mental deficiencies and perhaps some psychiatric disorders that are transmitted in a Mendelizing fashion, most of the attributes studied by soft-field psychologists are influenced by polygenic systems. Usually we must assume that several totally different and unrelated polygenic systems influence a manifest trait like social introversion. Introversion may be based in part on a unitary (although polygenic) variable, as shown by Gottesman (1963) and others. However, as an acquired disposition of the adult-acculturated individual, it presumably results from a confluence of different polygenic contributors such as basic anxiety readiness, mesomorphic toughness, garden-variety social introversion, dominance, need for affiliation, and the like.

6. *Divergent Causality*

As pointed out 35 years ago by the physical chemist Irving Langmuir (1943; London, 1946; Meehl, 1954, pp. 60–61; Meehl, 1967/1970b, especially Footnotes 1–8 on pp. 395–396), there are complex systems whose causal structure and boundary conditions are such that slight differences—including those that are, for practical predictive and explanatory purposes, effectively "random" (whatever their inner deterministic nature may be—tend to "wash out," "cancel each other," or "balance" over the long run. On the other hand, there are other systems in which such slight perturbations or differences in the exact character of the initial conditions are, so to speak, amplified over the long run. Langmuir christened the former kind of causality as "convergent," as when we say that the average errors in making repeated measurements of a table length tend to cancel out and leave us with a stable and highly trustworthy mean value of the result. On the other hand, an object in unstable equilibrium can lean slightly toward the right instead of the left, as a result of which a deadly avalanche occurs burying a whole village. Although both sorts of systems are found at all levels of Comte's Pyramid of the Sciences, it seems regrettably true that the incidence of important and pervasive types of divergent causality is greater in the sciences of behavior.

7. *Idiographic Problem*

It is not necessary to "settle" the long-continued methodological controversies regarding idiographic versus nomothetic methods in psychology and history (e.g., whether they are philosophically, metaphysically fundamentally different) to agree with strong proponents of the idiographic method, such as Gordon Allport (Allport, 1937) or my long-time friendly adversary on the prediction issue, Robert R. Holt (1958), that the human personality—unless one approaches it with the postulate of impoverished reality—has in its content, structure, and, conceivably, even in individual differences as to some of its "laws," and very much in its origins, properties and relations that make the study of personality rather more similar to such disciplines as history, archeology (historical), geology, or the reconstruction of a criminal case from police evidence than the derivation of the molar gas laws from the kinetic theory of heat or the mechanisms of heredity from molecular biology. Some would argue that such explanatory derivations aside, even the mere inductive subsumption of particulars (episodes, molar traits, persons) under descriptive generalizations is a more difficult and problematic affair in these disciplines than in most branches of physical and biological science.

8. *Unknown Critical Events*

Related to divergent causality and idiographic understanding but distinguishable from them is the fact that critical events in the history of personality development are frequently hard to ascertain. There is reason to believe that in some instances they are literally never ascertained by us or known to the individual under study, even somebody who has spent 500 hours on the analytic couch. They are sometimes observable events that, however, were not in fact observed and recorded, such as the precise tone of voice and facial expression that a patient's father had when he was reacting to an off-color joke that the patient innocently told at the dinner table at age 7. Every thoughtful clinician realizes that the standard life history that one finds in a medical chart is, from the standpoint of thorough causal comprehension, so thin and spotty and selective as to border on the ludicrous. But there is also what I would view as an important causal source of movement in one rather than another direction of divergent causality, namely, inner events, such as fantasies, resolutions, shifts in cognitive structure, that the patient may or may not report and that he or she may later be unable to recall.

9. *Nuisance Variables*

Other things equal, it is handy for research and theorizing if we can sort out the variables into three classes, namely, (a) variables that we manipulate (in the narrow sense of the word experimental), (b) variables that we do not manipulate but can hold constant or effectively exclude from influence by one or another means isolating the system under study, and (c) variables that are quasirandom with respect to the phenomena under study, so that they only contribute to measurement error or the standard deviation of a statistic. Unfortunately, there are systems, especially social and biological systems of the kind that clinical psychologists and personologists study, in which there is operative a nonnegligible class of variables that are not random but systematic, that exert a sizable influence, and are themselves also sizably influenced by other variables, either exogenous to the system (F. M. Fisher, 1966) or contained in it, such that we have to worry about the influence of these variables, but we cannot always ascertain the direction of the causal arrow. Sometimes we cannot even get sufficiently trustworthy measurements of these variables so as to "partial out" or "correct" their influence even if we are willing to make conjectures about the direction of causality. There are some circumstances in which we can extrapolate from experimental studies or from well-corroborated theory to make a high-confidence decision about the direction of causal influence, but there are many other circumstances—in soft psychology, the preponderating ones—in which this is not possible. Further, lacking special configurations such as highly atypical cells in a multivariate space or correlation coefficients that impose strong constraints on a causal interpretation, or provisional assumptions as relied on in path analysis (Li, 1975), the system is statistically and causally indeterminate. (Why these constraints are regularly treated as "assumptions" instead of refutable conjectures is itself a deep and fascinating question that I plan to examine some other time.) The well-known difficulties in assessing the influence of socioeconomic status (SES) on children's IQ when unscrambling the hereditary and environmental contributors to intelligence is perhaps the most dramatic one, but other less emotion-laden examples can be found on all sides in the behavioral sciences. (See Meehl, 1970a, 1971/1973b).

10. *Feedback Loops*

A special case in engineering is the usual in psychology, that a person's behavior affects the behavior of other persons and hence alters the schedule imposed by the "social Skinner box." The complexities here are so refractory to quantitative decomposition that yoked box setups came to be used even for the (relatively simple) animal case as a factual substitute for piecewise causal–dispositional analysis. In the human social case, they may be devastating.

11. *Autocatalytic Processes*

The chemist is familiar under the label *autocatalysis* with a rare but important kind of preparation in which one of the end products of the chemical processes is itself capable of catalyzing the process. Numerous common examples spring to mind in psychology, such as anxiety and depression as affects or economic failure as a social impact. Much of neurosis is autocatalytic in the cognitive-affective-volitional system, as are counterneurotic healing processes. When this kind of complicated setup is conjoined with the critical event idiographic, and divergent causality factors, and also with the individual differences factor (that parameters relating the growth of one state of schedule to a dependent variable, which itself in turn acts autocatalytically, show individual differences), the task of unscrambling such a situation becomes terribly difficult.

12. *Random Walk*

There is a widespread and understandable tendency to assume that the class of less-probable outcomes, given constancy of other classes of causally efficacious variables, should in principle be explicable by detecting a class of systematic input differences. Thus, for instance, we try to understand the genetic/environmental contributions to schizophrenia by studying discordant monozygotic twins. If I develop a florid clinical schizophrenia and my monozygotic twin remains sane and wins the Pulitzer Prize for poetry, it is a sensible strategy for the psychologist to consider my case *and similar cases* with an eye for "systematic differences" (such as who was born first, who was in what position in the uterus, or who had a severe case of scarlet fever with delirium) as responsible for dramatic difference in final outcomes. When one reflects on the rather meager yield of such assiduous ferreting out of systematic differences by, say Gottesman and Shields (1972) in their excellent book, one experiences bafflement. On the one hand, the concordance rate for monozygotic twins is only a little over 50%, indicating a very large nongenetic component in causality. Yet, on the other hand, we find feeble or null differences when we look at the list of "obvious, plausible" differentiators between the twins who fall ill and the twins who remain well. Of course, one can always say—and would no doubt be partly right in this—that we just have not been clever enough to hit on the right ones; or even if, qualitatively, they are the right ones, we do not have sufficiently construct-valid measures of them to show up in the statistics.

There is, however, an alternative explanation that when one reflects on it, is plausible (at least to a clinical practitioner like myself) and that has analogues in organic medicine and in other historical sciences like geology or the theory of evolution, to wit, that we are mistaken to look for a "big systematic variable" of the kind that is already in our standard list of influences, such as organic disease, parental preference, or SES of an adoptive home. Rather, we might emphasize that a human being's life history involves as one form of divergent causality, something akin to the stochastic process known as a "random walk" (Bartlett, 1955, pp. 15–20, 47–50, 89–96; Feller, 1957, pp. 73, 311; Kemeny, Snell, & Thompson, 1957, pp. 171–177; Read, 1972, pp. 779–782). At several points that are individually minor but collectively critical determinative, it is an almost "chance" affair whether the patient does A or not A, whether his girl friend says she will or will not go out with him on a certain evening, or whether he happens to hit it off with the ophthalmologist that he consults about some peculiar vision disturbances that are making him anxious about becoming blind, and the like. If one twin becomes psychotic at the end of such a random walk, it is possible that he was suffering from what was only, so to speak, "bad luck"—not a concept that appears in any standard list of biological and social nuisance variables!

Luck is one of the most important contributors to individual differences in human suffering, satisfaction, illness, achievement, and so forth, an embarrassingly "obvious" point that social scientists readily forget (Gunther, 1977; Jencks, 1972, pp. 8–9, 227–228; Popper, 1974, pp. 36–37; Stoddard, 1929; for further discussion of this see Meehl, 1972~~/1973g~~, pp. 402–407, Meehl, 1973d, pp. 220–221). Of course, the fact that a process resembles a random walk does not mean that it is not susceptible to quantitative treatment. Witness the extensive formal development of this sort of process in the field of finite mathematics by engineers and others. The point is that its analytical treatment will not look like the familiar kind of search for a systematic class of differentiating variables like SES as a nuisance variable in relationship to educational outcome and intelligence.

13. *Sheer Number of Variables*

I suppose that this is the most commonly mentioned of the difficulties of social science, and I assume that my readers would accept it without further elaboration. But it is worth mention that the number of variables is large from several different viewpoints. Thus we deal on one side with a *large number of phenotypic traits*, conceiving a phenotypic trait as a related family of response dispositions that (a) are correlated to some stipulated degree pairwise and that (b) have some kind of logical, semantic, social, or other "meaning" overlap or resemblance that entitles us to class them together. Or, again, we consider a large number of dimensions on the stimulus side and on the response side that are relevant in formulating a law of behavior acquisition, as well as in the subsequent control and activation dispositions thus acquired. From still another viewpoint, the list of historical causal influences is long and heterogeneous, ranging from such diverse factors as a mutated gene or a never-diagnosed subclinical tuberculosis to a mother who mysteriously absented herself the day after a patient first permitted himself the fantasy that a brutal father would go away, and the like. It should be noted that this matter of sheer number of variables would not be so important (except as a contributor to residual "random variation" in various kinds of outcomes) if they were each small contributors and independent, like the sources of error in the scattering of shots at a target in classical theory of errors. But in psychology this is riot typically the situation. Rather, the variables, although large in number, are each nuisance variables that carry a significant amount of weight, interact with each other, and contribute to idiographic development via the divergent causality mode.

14. *Importance of Cultural Factors*

This source of individual differences, both in acquired response clusters (traits) and in the parameters of acquisition and activation functions, especially when taken together with the genetic factors contributing, for instance, to social competence, mental health, intellect, and so on, makes for unusual complications in understanding how somebody got to be the way he is. We are, for instance, so accustomed to referring to nuisance variables like SES in considering the design of experiments that involve SES-related individual differences that we readily forget something every reflective person knows—that the measures of things like SES are general and not tailor-made for what is idiographically more significant in the development of a particular person. So when we speak of "controlling for SES," that is a loose use of language in comparison with "controlling the temperature" in a Skinner box or controlling the efflux of calories in a physics lab by use of a bomb calorimeter. A treatise on the principles of internal medicine (such as Harrison et al., 1966) sometimes refers to cultural factors, including those that are not at all understood—in the way that, say, dietary deficiency might be mediated by extreme poverty in a backward country—and simply says that for some reason this disease is found more frequently among the rich than among the poor. But the *important* causal chains of prime interest to the physician, even in his role as an advisor of preventive medicine, do not typically involve worry about whether somebody is fifth-generation upper class or the third child of parents who became anxious after the birth of the second oldest sibling. However, this kind of consideration might be crucial in reconstructing the life history of such a person.

15. *Context-Dependent Stochastotogicals*

Cronbach and Meehl (1955/1973) and subsequent writers adopted (from the neopositivist philosophers of science) the phrase *nomological network* to designate the system of lawlike relationships conjectured to hold between theoretical entities (states, structures, events, dispositions) and between theoretical entities and their observable indicators. The "network" metaphor is chosen to emphasize the structure of such systems, in which the *nodes* of the network, representing the postulated theoretical entities, are connected by the *strands* of the network, representing the lawful relationships hypothesized to hold between the entities. What makes such a set of theoretical statements a system (rather than a mere conjunction of unrelated assertions, a "heap of hypotheses") is the semantic fact of their shared terms, an overlap in the propositions' inner components, without which, of course, no deductive fertility and no derivation chains to observational statements would be formally possible. The network is empirical (and "scientifically respectable"), because a proper subset of the theoretical terms is coordinated in fairly direct ways ("operationally") with terms designating perceptual or instrument-reading predicates. These latter predicates normally possess the admirable properties of *quick decision, minimal theory dependence*, and *high interpersonal consensus*.

Despite the current distaste for these "objectivist" conceptions, I remain an old-fashioned unreconstructed positivist to the limited extent that I think science—both "normal science" and "revolutionary, paradigm-replacing science"—differs from less promising, noncumulative, and personalistic enterprises like politics, psychotherapy, folklore, ethics, metaphysics, aesthetics, and theology *in part* because of its skeptical insistence on reliable (intersubjective, replicable) protocols that describe observations. Skinner is in better shape than Freud partly because Norman Campbell (1920/1957, p. 29) was right in saying that the kinds of judgments for which universal assent can be obtained are (a) judgments of temporal simultaneity, consecutiveness, and "betweenness"; (b) judgments of coincidence and "betweenness" in space; and (c) judgments of number. I cannot view the increasingly fashionable dismissal of these objectivity-oriented views as other than obscurantist in tendency. (See Kordig, 1971, 1973.)

However, the nomological network, even though correlated directly, here and there, with observational data, is not "operational" throughout, since some of the nodes and strands are connected with the observational data base only via other subregions of the network. As Hempel said (1952):

> A scientific theory might therefore be likened to a complex spatial network: Its terms are represented by the knots, while the threads connecting the latter correspond, in part, to the definitions and, in part to the fundamental and derivative hypotheses included in the theory. The whole system floats, as it were, above the plane of observation and is anchored to it by rules of interpretation. These might be viewed as strings which are not part of the network but link certain points of the latter with specific places in the plane of observation. By virtue of those interpretive connections, the network can function as a scientific theory: From certain observational data, we may ascend, via an interpretive string, to some point in the theoretical network, thence proceed, via definitions and hypotheses, to other points, from which another interpretive string permits a descent to the plane of observation. (p. 36)

Even though the core of these ideas is sound and important, the word *nomological* is in soft psychology at best an extension of meaning and at worst a misleading corruption of the logician's terminology. Originally it designated strict laws as in W. E. Johnson's (1921/1964) earlier use of "nomic necessity" (p. 61). The lawlike relationships we have to work with in soft psychology are rarely (never?) of this strict kind, errors of measurement aside. Instead, they are correlations, tendencies, statistical clusterings, increments of probabilities, and altered stochastic dispositions. The ugly neologism *stochastological* (as analogue to *nomological*) is at least shorter than the usual "probabilistic relation" or "statistical dependence," so I shall adopt it. We are so accustomed to our immersion in a sea of stochastologicals that we may fail to notice what a terrible disadvantage this sort of probabilistic law network puts us under, both as to the clarity of our concepts and, more importantly, the testability of our theories. (One still hears the tiresome complaint that a theoretical system cannot be simultaneously concept definatory and factually assertive, despite repeated explanations of how this works. See, e.g., Braithwaite, 1960, pp. 76–87; Campbell, 1920/1957, pp. 119–158; Carnap, 1936–1937/1950,1952/1956, 1966, pp. 225–226, 265–274; Feigl, 1956, pp. 17–19; Hempel, 1952, 1958, pp. 81–87; Lewis, 1970; Maxwell, 1961, 1962; Meehl, 1977, pp. 35–37; Nagel, 1961, pp. 87, 91–93; Pap, 1958, pp. 318–321, 1962, pp. 46–52; Popper, 1974, pp. 14–73; Ramsey, 1931/1960; Sellars, 1948.)

When the observational corroborators of the theory consist wholly of percentages, crude curve fits, correlations, significance tests, and distribution overlaps, it is difficult or impossible to see clearly when a given batch of empirical data refutes a theory or even when two batches of data are (in any interesting sense) "inconsistent." All we can usually say with quasi-certainty is that context-dependent statistics should *not* be numerically identical in different studies of the same problem. (A dramatic recent example of this was the discovery that some of Sir Cyril Burt's correlation coefficients were *too consistent* to have been derived from the different tests and populations that he reported!)

In heading this section "Context-Dependent Stochastologicals," I mean to emphasize the aspect of this problem that seems to me most frustrating to our theoretical interests, namely, that the statistical dependencies we observe are always somewhat, and often strongly, dependent on the institution-cum-population setting in which the measurements were obtained. Lacking a "complete (causal) theory" of what influences what, *and how much*, we simply cannot compute expected numerical changes in stochastic dependencies when moving from one population or setting to another. Sometimes we cannot even rationally predict the direction of such changes. If the difference between two Pearson correlations were safely attributable to random sampling fluctuation alone, we could use the statistician's standard tools to decide whether Jones's study "fails to replicate" Smith's. But the usual situation is not one of simple cross-validation shrinkage (or "boostage")—rather, it involves the validity generalization problem. For this, there are no standard statistical procedures. We may be able, relying on strong theorems in general statistics plus a backlog of previous experience and a smattering of theory, to say some fairly safe things about restriction of range and the like. However, thoughtful theorists realize how little *quantitatively* we can say with sufficient confidence to warrant counting an unexpected shift in a stochastic quantity as a strong "discorroborator." This being so, we cannot fairly count an "in the ball park" predicted value as a strong corroborator. For example, Meehl's Mental Measure correlates .50 with SES in Duluth junior high school students, as predicted from Fisbee's theory of sociability. When Jones tries to replicate the finding on Chicano seniors In Tucson, he gets r = .34. Who can say anything theoretically cogent about this difference? Does any sane psychologist believe that one can do much more than shrug?

Although probability concepts (in the theory) and statistical distributions (in the data) sometimes appear in both classical and quantum physics, their usual role differs from that of context-dependent stochastologicals in social science. Without exceeding space limitations or my competence, let me briefly suggest some differences. When probabilities appear in physics and chemistry, they often drop out in the course of the derivation chain, yielding a quasi-nomological at its termination (e.g., derivation of gas laws or Graham's diffusion law from the kinetic theory of heat, in which the postulates are nomological, the "conditions" are probability distributions, and the resulting theorems are again nomological). Second, when the predicted observational result still contains statistical notions, their numerical values are either not context dependent or the context dependencies permit precise experimental manipulation. A statistical scatter function for photons or electrons can be finely tuned by altering a very limited number of experimental variables (e.g., wavelength, slit width, screen distance), and the law of large numbers assures that the expected "probabilistic" values of, say, photon incidence in a specified band will be indiscernibly different from the observed (finite but huge) numbers.

All this is very unlike the stochastologicals of soft psychology, in which strong context dependence prevails, but we do not know (a) the complete list of contextual influences, (b) the function form of context dependency for those influences that we can list, (c) the numerical values of parameters in those function forms that we know or guess, or (d) the values of the context variables if we are so fortunate as to get past Ignorances a–c. Finally, unlike physics, our sample sizes are usually such that the Bernoulli theorem does not guarantee a close fit between theoretical and observed frequencies—perhaps one of the few good uses for significance tests?

16. *Open Concepts*

As a consequence of the factors listed supra, the especially those numbered 4, 7, 9, 15, it is usually not possible in the soft areas of social science to provide rigorous, explicit, or—the holy word when I was in graduate school—operational definitions for theoretical concepts. This difficulty occurs not because psychologists are intellectually lazy or sloppy, although most of us are at times (some routinely and on principle). Rather, it arises from the intrinsic nature of the subject matter, that is, from the organism's real compositional nature and structure and the causal texture of its environment. As has often been pointed out, one can concoct quick and easy "operational definitions" of psychological terms, but they will usually lack theoretical interest and, except for some important special cases (e.g., purely predictive task-tailored psychometrics and some kinds of operant behavior control), generalizable technological power (Lazarus, 1971; Loevinger, 1957). It is remarkable evidence of cultural lag in intellectual life that one can still find quite a few psychologists who are hooked on the dire necessity of strictly operational definitions, and who view open concepts as somehow methodologically sinful, although it is now a quarter of a century since the late Arthur Pap published his brilliant article on open concepts (Pap, 1953, see also chap. 11 of Pap, 1958). To do justice, and highlight the cultural lag, I should mention the related article of Waismann that antedated Pap's by 8 years (Waismann, 1945) and even Carnap's of 40 years ago (1936–l937/1950). I cannot name a single logician or a philosopher (or historian) of science who today defends strict operationism in the sense that some psychologists claim to believe in it. (They don't really —but you have to listen awhile to catch the deviations in *substance* when pseudooperationists are not discoursing dogmatically about *method*.)

The problem of open concepts and their relation to empirical falsifiability warrants a separate article, with which I am currently engaged, but suffice it to say here that the unavoidability of open concepts in social and biological science tempts us to sidestep it by fake operationism on the one side (if we are of the tough-minded, superscientific orientation) or to be contented with fuzzy verbalisms on the other side (if we are more artsy-craftsy or literary), thinking that it is the best we can get. The important point for methodology of psychology is that just as in statistics one can have a *reasonably precise theory of probable inference*, being "quasi-exact about the inherently inexact," so psychologists should learn to be sophisticated and rigorous in their metathinking about open concepts at the substantive level. I do not mean to suggest in saying this that the logicians' theory of open concepts is in a highly developed state, but it is far more developed than one would think from reading or listening to most psychologists.

I have elsewhere (Meehl, 1977) distinguished three kinds of openness that are involved in varying degrees in various psychological concepts and that may all be present in the same theoretical construct, namely, (a) openness arising from the indefinite extensibility of our provisional list of operational indicators of the construct; (b) openness associated with each indicator singly, because of the empirical fact that indicators are only probabilistically, rather than nomologically, linked to the inferred theoretical construct; and (c) openness due to the fact that most of our theoretical entities are introduced by an implicit or contextual definition, that is, by their role in the accepted nomological network, rather than by their inner nature. By their "inner nature" I mean nothing spooky or metaphysical but merely their ontological structure or composition as the latter will, with the progress of research, be formulatable in terms of the theoretical entities of more basic sciences in Comte's pyramid. In social and biological science, one should keep in mind that *explicit definition* of theoretical entities is seldom achieved in terms of the initial observational variables of those sciences, but it becomes possible instead by theoretical reduction or fusion. Explicit definition is achieved, if ever, in terms of some more basic underlying science (Meehl, 1977, see also Cronbach & Meehl (1955/1973); Meehl, 1973f, 1973h, pp. 285–288).

A final remark, which also deserves fuller treatment in another place, is that when we deal with open concepts, as in personality psychometrics of traits or taxa, the statistical phenomenon of *psychometric drift* as a result of bootstrap operations, refinement of measures, and theoretical reflection on the big matrix of convergent and discriminative validities (Campbell & Fiske, 1959) also generates, via our reliance on implicit or contextual definitions of theoretical entities, an associated *conceptual drift*, a meaning shift. When we reassign weight to fallible indicators of an entity to the extent that the very meaning of the term designating that entity is specified by its role in the network, such reassignment of weights—especially under drastic revisions of the system such as dropping a previously relied-upon indicator—constitutes a change in the theoretical concept. Difficult interpretative and research strategy problems arise here, because, on the one hand (especially in psychometrics) we encounter the danger that the resulting conceptual drift has pulled us away from what we started out to measure, but we also recognize that in psychology, as in the other sciences, part of the research aim is precisely that of bringing about revisions of concepts on the basis of revisions of the nomological network that implicitly defines them. We want, as Plato said, to carve nature at its joints; and the best test of this achievement is increased order in our material.

17. *Intentionality, Purpose, and Meaning*

We do not need to settle the philosopher's question of what is the essential condition for the existence of intentionality, nor buy Brentano's famous criterion that intentionality is the distinctive mark of the mental, to recognize that human beings think and plan and intend, that if rats do so they do it at a much lower level, that sunflowers probably do not, and that stones certainly do not. The formulation of powerful functional relationships for systems that do not possess the capacity to think, worry, regret, plan, and intend is obviously on the average an easier task. (But see Vico, 1744/1948, for a view so different that an American social scientist of our time can hardly grasp it.)

18. *Rule Governance*

Related to intentionality but sufficiently important to deserve a special listing is the fact that human behavior is rule governed. People do something not merely "in accordance with" a generalization but because they feel bound to obey the generalization stated in the form of a rule. Nobody has succeeded in coming up with a fully satisfactory definition of when a rule is a rule, but a sufficiently good approximation is to say that a rule differs from an empirical generalization in that a rule is not liquidated by being broken, whereas an empirical generalization is thereby liquidated (assuming that the conditions stated in its antecedent clause are granted, and the violation event is admitted into the corpus). Continued controversies in psycholinguistics reflect the importance of this kind of consideration in any discussion of human conduct.

19. *Uniquely Human Events and Powers*

In addition to being rule governed, there are several other human features that we do not share with chimpanzees, let alone sponges or boulders. I recall the late Richard M. Elliott saying that the main reason that psychology had done so poorly in its "theories" of humor is that man is the only animal that laughs. I think he had a good point here, since we have learned so much about aspects of human functioning, such as digestion and reproduction, by the experimental study of animals. There are a number of other things that human beings do that no infrahuman animal does, so far as we know. Only man speculates about nonpractical, theoretical matters; only man worships; only man systematically goes about seeking revenge, years later, for an in jury done to him; only man carries on discussions about how to make decisions; and there are some features of cultural transmission that only man engages in, although the evidence now indicates that numerous other species transmit learned forms of behavior to subsequent generations.

20. *Ethical Constraints on Research*

This one is so obvious as to need no exposition. One can readily conceive quasi-definitive experiments on the IQ–heredity controversy, or whether there are family dynamics sufficient to make just anyone into a manic-depressive, that cannot be performed because to do so would be immoral.

Not to be overly pessimistic, let me mention (without proof) five noble traditions in clinical psychology that I believe have permanent merit and will still be with us 50 or 100 years from now, despite the usual changes. Some of these are currently unpopular among those addicted to one of the contemporary fly-by-night theories, but that does not bother me. These five noble traditions are (a) descriptive clinical psychiatry, (b) psychometric assessment, (c) behavior genetics, (d) behavior modification (I lump under this rubric positive contingency management, aversion therapy, and desensitization), and (e) psychodynamics. This list should convince you that I am not using methodological arguments to grind any substantive ax. I am probably one of the few psychologists alive today who would list all five of these as great, noble, and enduring intellectual traditions. I particularly emphasize the last, psychodynamics, since I am often perceived as a dust bowl empiricist who does not think that anything can be true or useful if it is not either based on laboratory experiments or statistical correlations. There is not a single experiment reported in my 23-volume set of the standard edition of Freud nor is there a *t* test. But I would take Freud's clinical observations over most people's *t* tests any time. I am confident that psychoanalytic concepts will be around after rubber band theory, transactional theory, attachment theory, labeling theory, dissonance theory, attribution theory, and so on, have subsided into a state of innocuous desuetude like risky shift and level of aspiration. At the very least, psychoanalysis is an interesting theory, which is more than I can say about some of the "theories" that are currently fashionable.

These five noble traditions differ greatly in the methods they use and their central concepts, and I am hard put to say what is common among them. Some of them, such as behavior modification, are not conceptually exciting to those of us who are interested in ideas like Freud's, but they more than make up for that by their remarkable technological power.

I shall focus the remainder of my remarks on one feature that they have in common with the developed sciences (physical or biological); to wit, they were originally developed with negligible reliance on *statistical significance testing*. Even the psychometric assessment tradition in its early stages paid little attention to significance testing except (some times) for finding good items. Binet did not know anything about *t* tests, but he drew graphs of the developmental change of items. I suggest to you that Sir Ronald has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology.

It is easiest to see this from the methodological viewpoint of Sir Karl Popper, but fortunately we have here a rare instance in which Sir Karl's position yields the same result as the Bayesians', and both give the same result as "scientific common sense" practiced by those chemists and biologists who know nothing about philosophy of science or Bayesian statistics and could not care less about either. Briefly and simplistically, the position of Popper and the neo-Popperians is that we do not "induce" scientific theories by some kind of straightforward upward seepage from the clearly observed facts, nor do we "confirm" theories as the Vienna positivists supposed. All we can do is to subject theories—including the wildest and "unsupported" armchair conjectures (for a Popperian, completely kosher) —to grave danger of refutation, in accordance with the formally valid fourth figure of the implicative syllogism: $p \rightarrow q, \sim q, \therefore \sim p$, Popper's famous *modus tollens*.

A theory is corroborated to the extent that we have subjected it to such risky tests; the more dangerous tests it has survived, the better corroborated it is. If I tell you that Meehl's theory of climate predicts that it will rain sometime next April, and this turns out to be the case, you will not be much impressed with my "predictive success." Nor will you be impressed if I predict more rain in April than in May, even showing three asterisks (for $p < .001$) in my *t*-test table! If I predict from my theory that it will rain on 7 of the 30 days of April, and it rains on exactly 7, you might perk up your ears a bit, but still you would be inclined to think of this as a "lucky coincidence." But suppose that I specify *which* 7 days in April it will rain and ring the bell; then you will start getting seriously interested in Meehl's meteorological conjectures. Finally, if I tell you that on April 4th it will rain 1.7 inches (.66 cm), and on April 9th, 2.3 inches (.90 cm) and so forth, and get seven of these correct within reasonable tolerance, you will begin to think that Meehl's theory must have a lot going for it. You may believe that Meehl's theory of the weather, like all theories, is, when taken literally, false, since probably all theories are false in the eyes of God, but you will at least say, to use Popper's language, that it is beginning to look as if Meehl's theory has considerable *verisimilitude*, that is, "truth-likeness." (An adequate reconstruction of the verisimilitude concept has yet to be provided by our logician friends, see, e.g., Popper, 1976, but few reflective psychologists will doubt that some such notion of "nearness to the truth" is unavoidable when we evaluate theories. It is crucial to recognize that verisimilitude is an ontological, not an epistemological, concept that must not be conflated with confirmation probability, evidence, proof, corroboration, belief, support, or plausibility.)

Popperians would speak of low logical or prior probability, of the high content (forbidding much), because it specifies exactly which days it will rain how many inches. A Bayesian (who would reject Popper's philosophy on the grounds that we want our "theoretical prior" to be *high* to get a nice boost out of Bayes' theorem when the facts turn out right) would express Popper's point by saying that we want what Pap (1962, p. 160) calls the *expectedness*, the prior on the observations that is found in the denominator of Bayes' theorem to be low. An unphilosophical chemist or astronomer or molecular biologist would say that this was just good sensible scientific practice, that a theory that makes precise predictions and correctly picks out *narrow intervals* or *point values* out of the range of experimental possibilities is a pretty strong theory. There are revisions (as I think, necessary) of the classic Popperian position urged on us by his heretical exstudents P. K. Feyerabend and the late Imre Lakatos, but psychologists must reach at least the stage of Bayes and Popper before they can profitably go on to the refinements and criticisms of these gentlemen.

The most important caveat I would adjoin to Sir Karl's falsifiability requirement arises from the considerations pressed by Feyerabend (1962, 1965, 1970, 1971), Lakatos (1970, 1974a, 1974b), and others concerning the crucial role of auxiliary theories in subjecting the main substantive theory of interest to danger of *modus tollens*. As is well-known (and not disputed by Popper), when we spell out in detail the logical structure of what purports to be an observational test of a theoretical conjecture T, we normally find that we cannot get to an observational statement from T alone. We require further a set of often complex and problematic auxiliaries A, plus the empirical realization of certain conditions describing the experimental particulars, commonly labeled collectively as C. So that the derivation of an observation from a substantive theory T amounts always to the longer formula $(T.A.C) \rightarrow O$, rather than the simplified schema $(T \rightarrow O)$ that most of us learned in undergraduate logic courses. This presents a problem not perhaps for Popper's main thesis (although some critics do say this) but for its application as a criterion of the scientific status of theories (or the scientific approach of a particular theoretician or investigator?). The *modus tollens* now reads: Since $(T.A.C) \rightarrow O$, and we have falsified O observationally, we have the consequence $\sim(T.A.C)$. Unfortunately, this result does not entail the falsity of T, the substantive theory of interest but only the falsity of the conjunction $(T.A.C)$; that is, we have proved a disjunction of the falsities of the conjuncts. So the failure to get the expected observation O proves that $\sim T \vee \sim A \vee \sim C$, which is not quite what we would like to show.

One need not subscribe to the famous Duhemian thesis regarding falsification of science as a whole (Grünbaum, 1960, 1962, 1969, 1976) or to the Lakatosian exposition (Lakatos, 1970, 1974a, 1974b) about the protecttive belt of auxiliaries against which the *modus tollens* is directed versus the hard core of the theory against which the *modus tollens* is, prior to a Kuhnian revolution (Kuhn, 1970a, 1970b, 1970c), forbidden to be directed, to see that there is a difficult problem presented to even a neo-Popperian (like myself), because in social science the auxiliaries *A* and the initial and boundary conditions of the system *C* are frequently as problematic as the theory *T* itself. *Example*: Suppose that a personologist or social psychologist wants to investigate the effect of social fear on visual perception. He attempts to mobilize anxiety in a sample of adolescent males, chosen by their scores on the Social Introversion (*Si*) scale of the Minnesota Multiphasic Personality Inventory (MMPI), by employing a research assistant who is a raving beauty, instructing her to wear Chanel No. 5, and adopt a mixed seductive and castrative manner toward the subjects. An interpretation of a negative empirical result leaves us wondering whether the main substantive theory of interest concerning social fear and visual perception has been falsified, or whether only the auxiliary theories that the *Si* scale is valid for social introversion and that attractive but hostile female experimenters elicit social fear in introverted young males have been falsified. Or perhaps even the particular conditions were not met; that is, she did not consistently act the way she was instructed to or the MMPI protocols were misscored.

There is nothing qualitatively unique about this problem for the inexact sciences, but it is quantitatively more severe for us than for the chemist or astronomer, for at least two reasons, which I shall set forth without either proving or developing them here. First, in dependent testing of the auxiliary theories (which often means validation of psychometric instruments or ascertaining efficacy of social stimulus inputs) is harder to carry out. Due to unavoidable looseness of the nomological network (Cronbach & Meehl, 1955/1973) plus the factors in the list of 20 difficulties supra, the range of research circumstances in which auxiliaries *A* are problematic is greater than in the exact sciences or in some but not all of the biological sciences. Second, a point to which philosophers of science have devoted little attention, in physics or chemistry there is usually a more intimate connection, sometimes one of contributing to derivability, between the substantive theory of interest *T* and components of the auxiliaries *A*. This is sometimes even true in advanced branches of biology. *Example*: There is a complicated, well-developed, and highly corroborated theory of how a cyclotron works, and the subject matter of that auxiliary "theory of the instrument" is for the most part identical to the subject matter of the physical theories concerning nuclear particles, and so on, being investigated by the physicist. Devices for bringing about a state of affairs, for isolating the system under study, and for observing what occurs as a result are all themselves legitimated by theory.

It seems there is a sense in which auxiliary theories used by physical and biological scientists are at least subtly informed by what may be loosely called *the spirit*, the leading ideas, the core, and pervasive concepts of the main substantive theory *T*, although not rigorously derivable from *T*. When this is not so, scientists are likely to consider the (*T.A.*) system as "unaesthetic," "incoherent," even ad hoc. These fascinating matters remain to be analyzed and reconstructed by logicians, but most scientists and historians of science are—however informally—well aware of their influence. (See, e.g., Holton, 1973.)

In the social sciences, no such intimate connection, and almost never a relation of theoretical derivability, exists; hence, the auxiliary theory (such as a theory that the Rorschach is valid for detecting subclinical schizoid cognitive slippage or that Chanel-doused beauteous research assistants are anxiety elicitors) must stand on its own feet. Almost nothing we know or conjecture about the substantive theory helps us to any appreciable degree in firming up our reliance on the auxiliary. The situation in which *A* is merely conjoined to *T* in setting up our test of *T* makes it hard for us social scientists to fulfill a Popperian falsifiability requirement—to state before the fact what would count as a strong falsifier.

I shall illustrate this problem further with a simple example whose adequate exposition will appear elsewhere (Golden & Meehl, in press). Suppose that I wish to test my dominant gene conjecture (Golden & Meehl, 1978; Meehl, 1972, 1972/1973g, 1977) concerning *schizotaxia* as the central nervous system condition for the development by social learning of *schizotypy* (Meehl, 1962/1973c), which in turn is the personality precondition for the development of a *clinical schizophrenia*—although the latter must then occur only in one fourth of the persons carrying the gene, given the roughly 12% concordance for first-degree relatives as regards diagnosable clinical schizophrenia. (See also Böök, 1960; Heston, 1966, 1970; Slater, 1958/1971). I might rely on some complex neurological or projective or structured test "sign" as having such-and-such estimated construct validity for the schizotypal personality makeup. Such a quantitative estimate might be made relying on a combination of empirical evidence concerning discordant monozygotic twins of known schizophrenics, protocols of persons tested as college freshmen who subsequently decompensate into a recognizable schizophrenia, and the like. Such numerical estimates will all suffer not only from the usual test unreliability and random sampling fluctuations, but they will also have some unknown degree of systematic bias. For instance, it clearly will not do to assume that the taxon *all compensated schizotypes* would average the same scores on a Rorschach or MMPI indicator variable as do the compensated (discordant) monozygotic twins, the latter being a biased selection, since they have the same potentiating genes that their decompensated twins have. However, there must be something else about them—of an environmental sort—that works strongly in their favor and helps keep them discordant, that is, clinically well. One simply has no way of ascertaining the net impact of these two opposed kinds of forces on the psychometric results.

Suppose that we take some combination of earlier findings on preschizophrenics, remitted schizophrenics, compensated discordant monozygotic twins of schizophrenics, and so forth, and we ascertain that while the valid positive rate p_s, among these safely presumed schizotypes varies (even if the sample sizes are huge, it will always vary in an amount unexplainable by random sampling fluctuation), it nevertheless shows a "reasonably close" agreement. (Again, we think like physicists or physiologists instead of like social scientists fooling around with *t* tests.) So we strike some kind of rough average $\overline{p}_s$ of these several valid positive rates, knowing that it is the best we can do at this point with data on different groups of schizotypes, who, despite their differences, must all have somehow been tagged as such. Given that estimated valid positive rate, and given a false positive rate p_n (also systematically biased because of the undiagnosed compensated schizotypes in any "control population"), we record our numerical predictions for the incidence of our psychometric sign among parent pairs of schizophrenic probands (where, on the dominant gene theory, we expect not only a 50% schizotypy incidence but something stronger; to wit, at least one member of each parent pair must be a schizotype). We also compute it for siblings and dizygotic twins and—although here things get a bit feeble—with sufficiently large samples, maybe second-degree relatives. Thus, for instance, the expected sign-positive rate among parents (and sibs, if they all cooperate) is given by the simple expression $p^+ = (1/2)\,p_s + (1/2)\,p_n$.

Now the substantive dominant gene theory *T*, when conjoined with the auxiliary theory *A* concerning psychometric validity, and assuming that we have identified the right relatives and the probands were all schizophrenics [=*C*], generates point predictions and therefore takes a high Popperian risk *when the conjunction* (*T.A.C*) *is considered as the "theory" under test*. Hence, the verification of those numerical point predictions as to the values of the psychometric incidence in relatives of different degrees of consanguinity provides a strong Popperian test for that conjunctive "theory." One would then normally say that successful negotiation of this hurdle, the failure to be clobbered *modus tollens* by the outcome of the empirical study, provides a moderate to strong corroboration of the conjunctive theory. Hence, (*T.A.C*) is doing well; that is, it has escaped falsification despite taking a high risk by making several numerical point predictions.

So far, so good, and Popper as well as his critics would have no complaint. However, the classical Popperian requirement on playing the scientific game fairly involves the theoretician's saying, before doing the research, what would count as a strong basis for rejecting the theory. If "the theory" is taken to be the substantive theory *T* (which it is, if one is not being philosophically disingenuous) rather than the psychometric auxiliary and diagnostic validity conjectures *A* and *C*, then one will be committing what amounts in spirit to a Popperian sin against falsificationism as a method. If the empirical research does not pan out as predicted, one does not abandon *T*; instead he tells us that either *T* is incorrect, *A* is incorrect, or the diagnoses were untrustworthy!

I am not persuaded from his writings nor from conversations that I have had with him that Sir Karl adequately appreciates the degree to which this theory and auxiliary problem permeate research in the inexact sciences, especially the social sciences in their soft areas. Whether it presents a general problem for the Popperian formulation of scientific method is beyond the scope of this article and my competence. It is perhaps worth saying, however, for the benefit of philosophically oriented readers, that the above described situation—certainly no rarity in our field or in biology—may represent a social fact about the way science works that presents grave difficulties for the Popperian reconstruction. That is, the stipulation beforehand that one will be pleased about substantive theory *T* when the numerical results come out as forecast, but will not necessarily abandon it when they do not, seems on the face of it to be about as blatant a violation of the Popperian commandment as you could commit. For the investigator, in a way, is doing what Popper says we ought not to do, and what astrologers and Marxists and psychoanalysts allegedly do, playing "heads I win, tails you lose." But it seems in accordance with much scientific practice and, as far as I have sampled, with most persons' scientific common sense or intuitions, to say that if the combination (*T.A.C*) generates a high-risk numerical point prediction, such a result really does support all three of the components. The reason it does so seems pretty clear, despite its commonsense, non-formalized character: Because of the lack of intimate inner connection in the inexact sciences between the components of these conjunctions, it would strike us as a very strange coincidence if the substantive theory *T* should have low verisimilitude (which would, were *T* true, also generate mispredictions of the numerical point values) and yet the two (largely unrelated) "wrongs" of *T* and *A* are somehow systematically balanced so as to generate the same numerical prediction generated from the conjecture that *T* and *A* both have relatively high verisimilitude.

Such a delicate quantitative counterbalancing of theoretical errors is not impossible, but it seems quite implausible, assuming that nature is (as Einstein says) "subtle but not malicious." So I think we are not being unreasonable to congratulate ourselves on arriving at a successful prediction of high-risk point values or other antecedently improbable observational patterns from the conjunction (*T.A.C*), despite the fact that we seem to be hedging when we say before the fact that we will not consider our substantive theory *T* falsified by a bad result if it does not pan out. These are problems that need further exploration by statisticians and philosophers of science, especially in light of work on the history of science, and with special attention to the question of whether there are important differences between the inexact and the exact sciences, or even between the biological and social sciences, as to how a Popperian or neo-Popperian methodology should be explained and applied.

But, you may say, what has all this got to do with significance testing? Isn't the social scientist's use of the null hypothesis simply the application of Popperian (or Bayesian) thinking in contexts in which probability plays such a big role? No, it is not. One reason it is not is that the usual use of null hypothesis testing in soft psychology as a means of "corroborating" substantive theories does not subject the theory to grave risk of refutation *modus tollens*, but only to a rather feeble danger. The kinds of theories and the kinds of theoretical risks to which we put them in soft psychology when we use significance testing as our method are *not* like testing Meehl's theory of weather by seeing how well it forecasts the number of inches it will rain on certain days. Instead, they are depressingly close to testing the theory by seeing whether it rains in April at all, or rains several days in April, or rains in April more than in May. It happens mainly because, as I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false. I shall not attempt to document this here, because among sophisticated persons it is taken for granted. (See Morrison & Henkel, 1970, especially the chapters by Bakan, Hogben, Lykken, Meehl, and Rozeboom.) A little reflection shows us why it has to be the case, since an output variable such as adult IQ, or academic achievement, or effectiveness at communication, or whatever, will always, in the social sciences, be a function of a sizable but finite number of factors. (The smallest contributions may be considered as essentially a random variance term.) In order for two groups (males and females, or whites and blacks, or manic depressives and schizophrenics, or Republicans and Democrats) to be *exactly* equal on such an output variable, we have to imagine that they are exactly equal *or* delicately counterbalanced on all of the contributors in the causal equation, which will never be the case.

Following the general line of reasoning (presented by myself and several others over the last decade), from the fact that the null hypothesis is always false in soft psychology, it follows that the probability of refuting it depends wholly on the sensitivity of the experiment—its logical design, the net (attenuated) construct validity of the measures, and, most importantly, the sample size, which determines where we are on the statistical power function. Putting it crudely, if you have enough cases and your measures are not to-tally unreliable, the null hypothesis will always be falsified, *regardless of the truth of the substantive theory*. Of course, it could be falsified in the wrong direction, which means that as the power improves, the probability of a corroborative result approaches one-half. However, if the theory has no verisimilitude—such that we can imagine, so to speak, picking our empirical results randomly out of a directional hat apart from any theory—the probability of refuting by getting a significant difference in the wrong direction also approaches one-half. Obviously, this is quite unlike the situation desired from either a Bayesian, a Popperian, or a commonsense scientific standpoint. As I have pointed out elsewhere (Meehl, 1967/1970b; but see criticism by Oakes, 1975; Keuth, 1973; and rebuttal by Swoyer & Monson, 1975), an improvement in instrumentation or other sources of experimental accuracy tends, in physics or astronomy or chemistry or genetics, to subject the theory to a greater risk of refutation *modus tollens*, whereas improved precision in null hypothesis testing usually decreases this risk. A successful significance test of a substantive theory in soft psychology provides a feeble corroboration of the theory because the procedure has subjected the theory to a feeble risk.

But, you may say, we do not look at just one; we look at a batch of them. Yes, we do; and how do we usually do it? In the typical *Psychological Bulletin* article reviewing research on some theory, we see a table showing with asterisks (hence, my title) whether this or that experimenter found a difference in the expected direction at the .05 (one asterisk), .01 (two asterisks!), or .001 (three asterisks!!) levels of significance. Typically, of course, some of them come out favorable and some of them come out unfavorable. What does the reviewer usually do? He goes through what is from the standpoint of the logician an almost meaningless exercise; to wit, he *counts noses*. If, say, Fisbee's theory of the mind has a batting average of 7:3 on 10 significance tests in the table, he concludes that Fisbee's theory seems to be rather well supported, "although further research is needed to explain the discrepancies." This is scientifically a preposterous way to reason. It completely neglects the crucial asymmetry between confirmation, which involves an inference in the formally invalid third figure of the implicative syllogism (this is why inductive inferences are ampliative and dangerous and why we can be objectively wrong even though we proceed correctly), and refutation, which is in the valid fourth figure, and which gives the *modus tollens* its privileged position in inductive inference. Thus the adverse *t* tests, seen properly, do Fisbee's theory far more damage than the favorable ones do it good.

I am not making some nit-picking statistician's correction. I am saying that the whole business is so radically defective as to be scientifically almost pointless. This is not a technical hassle about whether Fisbee should have used the varimax rotation, or how he estimated the communalities, or that perhaps some of the higher order interactions that are marginally significant should have been lumped together as a part of the error term, or that the covariance matrices were not quite homogeneous. I am not a statistician, and I am not making a statistical complaint. I am making a philosophical complaint or, if you prefer, a complaint in the domain of scientific method. I suggest that when a reviewer tries to "make theoretical sense" out of such a table of favorable and adverse significance test results, what the reviewer is actually engaged in, willy-nilly or unwittingly, is meaningless substantive constructions on the properties of the statistical power function, and almost nothing else.

This feckless activity is made worse by the almost universal practice of what I call *stepwise low validation*. By this I mean that we rely on one investigation to "validate" a particular instrument and some other study to validate another instrument, and then we correlate the two instruments and claim to have validated the substantive theory. I do not argue that this is a scientific nothing, but it is about as close to a nothing as you can get without intending to. Consider that I first show that Meehl's Mental Measure has a validity coefficient (against the criterion I shall here for simplicity take to be quasi-infallible or definitive) of, say, .40—somewhat higher than we usually get in personology and social psychology! Then I show that Glotz's Global Gauge has a validity for its alleged variable of the same amount. Relying on these results, having stated the coefficient and gleefully recorded the asterisks showing that these coefficients are not zero (!), I now try to corroborate the Glotz-Meehl theory of personality by showing that the two instruments, each having been duly "validated," correlate .40, providing, happily, some more asterisks in the table. Now just what kind of a business is this? Let us suppose that each instrument has a reliability of .90 to make it easy. That means that the portion of construct-valid variance for each of the devices is around one fifth of the reliable variance and the same for their over-lap when correlated with each other. I do not want to push the discredited (although recently revived) principle of indifference, but without other knowledge, it is easily possible, and one could perhaps say rather likely, that the correlation between the two occurs in a region of each one's components that has literally nothing to do with either of the two criterion variables used in the validity studies relied on. This is, of course, especially dangerous in light of the research that we have on the contribution of methods variance.

I seem to have trouble conveying to my students and colleagues just how dreadful a mess of flabby inferences this kind of thing involves. It is as if we were interested in the effect of sunlight on the mating behavior of birds, but not being able to get directly at either of these two things, we settle for correlating a proxy variable like field-mice density (because the birds tend to destroy the field mice) with, say, incidence of human skin cancer (since you can get that by spending too much time in the sun!). You may think this analogy dreadfully unfair; but I think it is a good one. Of course, the whole idea of simply counting noses is wrong, because a theory that has seven facts for it and three facts against it is *not* in good shape, and it would not be considered so in any developed science.

You may say, "But, Meehl, R. A. Fisher was a genius, and we all know how valuable his stuff has been in agronomy. Why shouldn't it work for soft psychology?" Well, I am not intimidated by Fisher's genius, because my complaint is not in the field of mathematical statistics, and as regards inductive logic and philosophy of science, it is well-known that Sir Ronald permitted himself a great deal of dogmatism. I remember my amazement when the late Rudolf Carnap said to me, the first time I met him, "But, of course, on this subject Fisher is just mistaken: surely you must know that." My statistician friends tell me that it is not clear just how useful the significance test has been in biological science either, but I set that aside as beyond my competence to discuss. The shortest answer to this rebuttal about agronomy, and one that has general importance in thinking about soft psychology, is that we must carefully distinguish *substantive theory* from *statistical hypothesis*. There is a tendency in the social sciences to conflate these in talking about our inferences. (A neglected article by Bolles, 1962, did not cure the psychologists' disease.) The substantive theory is the theory about the causal structure of the world, the entities and processes underlying the phenomena; the statistical hypothesis is a much more restricted and "operational" conjecture about the value of some parameter, such as the mean of a specified statistical population. The main point in agronomy is that the logical distance, the difference in meaning or content, so to say, between the alternative hypothesis and substantive theory *T* is so small that only a logician would be concerned to distinguish them. *Example*: I want to find out whether I should be putting potash on the ground to help me raise more corn. Now everybody knows from common sense as well as biology that the corn gets its nutrients from the soil, and furthermore that the yield of corn at harvest time is not causally efficacious in determining what I did in the spring, random numbers aside. If I refute the statistical null hypothesis that plots of corn with potash do not differ in yield from plots without potash, I have thereby proved the alternative hypothesis—that there *is* a difference between these two sorts of plots; and the only substantive conclusion to draw, given such a difference, is that the potash made the difference. Such a situation, in which the content of the substantive theory is logically quasi-identical with the alternative hypothesis, which was refuted by our significance test, is completely different from the situation in soft psychology. Fisbee's substantive theory of the mind is not equivalent, or anywhere near equivalent, to the alternative hypothesis. All sorts of competing theories are around, including my grandmother's common sense, to explain the nonnull statistical difference. So the psychologist can take little reassurance about the use of significance tests from knowing that Fisher's approach has been useful in studying the effect of fertilizer on crop yields.

Although this presents a pretty depressing picture, I daresay that the Skinner disciples among you will be inclined to think,

> well, that's just one more way of showing what we have known all along. The point is to prove that you have achieved experimental control over your subject matter, as Skinner says. If you have, I am not much interested in tabular asterisks; if you haven't, I'm not interested in them either.

But that is easy for Skinnerians because their theory (it is a theory in Sir Karl Popper's sense) is close to a pure dispositional theory and does not usually present us with the kind of evidentiary evaluation problem that we get with entity-postulating theories such as those of Freud, Hull, Albert Ellis, or, to come closer to home, my conjectures about schizophrenia or hedonic deficit (Meehl, 1972, 1974, 1975, 1962/1973c, 1972/1973g). Those of us whose cognitive passions are incompletely satisfied by dispositional theories, whether Skinnerian or psychometric, should ask ourselves what kind of inferred entity construction we want and how it could generate the sorts of intellectual "surprises" that Robert Nozick (1974, pp. 18–22) considers typical of invisible hand theories, which have proved so eminently successful in the physical and biological sciences and—somewhat less so—in economics. Some directions of solution (before I go on to the one that I am using in my own research) follow.

We could take the complex form of Bayes's theorem more seriously in concrete application to various substantive theories to take into account, even if crudely in the sense of setting upper and lower hounds to the probabilities involved, the logical asymmetry between confirmation and refutation (see, e.g., Maxwell, 1974). Second, it may be that the Fisherian tradition, with its soothing illusion of quantitative rigor, has inhibited our search for stronger tests, so we have thrown in the sponge and abandoned hope of concocting substantive theories that will generate stronger consequences than merely "the Xs differ from the Ys." Thus, for instance, even when we cannot generate numerical point predictions (the ideal case found in the exact sciences), it may be that we can at least predict the order of numerical values or the rank order of the first-order numerical differences, and the like.

Sometimes in the other sciences it has been possible to concoct a middling weak theory that, while incapable of generating numerical point values, entails a certain *function form*, such as a graph should be an ogive or that it should have three peaks and that these peaks should be increasingly high, and that the distance on the abscissa between the first two peaks should be less than the distance between the second two. In the early history of quantum theory, physicists relied on Wien's law, which related "some (unknown) function" of wavelength to energy multiplied by the fifth power of wavelength. In the cavity radiation experiment, the empirical points were simply plotted at varying temperatures, and it was evident by inspection that they fell on the same curve, even though a formal expression for that curve was beyond the theory's capabilities (Eisberg, 1961, pp. 5–5l).

Talking of Wien's law is a good time for me to recommend to psychologists who disagree with my position to have a look at any textbook of theoretical chemistry or physics, where one searches in vain for a statistical significance test (and finds few confidence intervals). The power of the physicist does not come from exact assessment of probabilities that a difference exists (which physicists would view as a ludicrous thing to show), nor by the verbal precision of so-called "operational definitions" in the embedding text. The physicist's scientific power comes from two other sources, namely, the immense deductive fertility of the formalism and the accuracy of the measuring instruments. The scientific trick lies in conjoining rich mathematics and experimental precision, a sort of "invisible hand wielding fine calipers." The embedding text is sometimes surprisingly loose, free-wheeling, even metaphorical—as viewers of television's *Nova* are aware, seeing Nobel laureates discourse whimsically about the charm, strangeness, and gluons of nuclear particles (see, e.g., Nambu, 1976). One gets the impression that when you have a good science going, with potent mathematics and accurate instruments, you can be relaxed and easygoing about the words. Nothing is as stuffy and pretentious as the verbal "pseudorigor" of the soft branches of social science. In my modern physics text, I am unable to find one single test of statistical significance. What happens instead is that the physicist has a sufficiently powerful invisible hand theory that enables him to generate an expected curve for his experimental results. He plots the observed points, looks at the agreement, and comments that "the results are in reasonably good accord with theory." Moral: *It is always more valuable to show approximate agreement of observations with a theoretically predicted numerical point value, rank order, or function form, than it is to compute a "precise probability" that something merely differs from something else*. Of course, we do not have precise probabilities when we do significance testing because of the falsity of the assumptions generating the table's values and varying robustness of our tests under departures from these assumptions.

The only possible "solution" to the theory-refutation problem that I have time to discuss in any detail is what I call *consistency tests* (Meehl, Note 3). Unfortunately, this approach is not easily available for most theoretical problems in soft psychology, although I am not prepared to say that it is confined to the domain in which I have been developing it, namely, taxometrics, that is, the application of psychometric procedures to detection of a taxonic situation and classification of individuals into the taxon or outside of it. From our conjectures about the latent causal situation, we derive formulas for estimating the theoretical quantities of interest, such as the proportion of schizotypes in a given clinical population, the mean values of the schizotypal and nonschizotypal classes, the optimal cut ("hitmax") on each phenotypic indicator variable for classifying individuals, and the proportion of valid and false positives achieved by that cut. But we realize that our conjectures about the latent situation may be false or that the indicators relied on may have too low validity, or that they may be more correlated within the taxa than desired, and so forth. Second, even if the basic formal structure postulated is approximated by the state of nature (e.g., there is a schizoid taxon, the indicators have sizable validity, the intrataxon distributions are quasi-normal or at least unimodal, the correlation of the indicators within the groups is small, and the departures from these various hypotheses are within the tolerance allowed by the method's robustness), it may still be that we have suffered some kind of systematic bias on one of the indicators due to a nuisance variable such as social class, or that we have had bad luck in the sample, so the method's numerical deliverances on this occasion are untrustworthy.

Whether the abstract causal structure postulated is unsound or the numerical values found in this sample are seriously in error, we need some method of checking the data internally to find out whether these unfortunately possibilities have materialized. We do this by deriving theorems within the formalism specifying how various numerical values (observed or calculated from the observed) should be related to each other, so that when they are not related as the consistency theorem demands, we are alerted to the danger that something is rotten in the state of Denmark (see Meehl, 1973d). Unfortunately, most of the work, both mathematical and empirical, is as yet only available in mimeographed reports from our laboratory (Golden, 1976; Golden & Meehl, Note 1, Note 2; Meehl, Note 3, Note 4). What survives scrutiny will be found in a book in preparation with my former student and research colleague Robert Golden (Golden & Meehl, in press).

One taxometric procedure, which I have christened *maxcov-hitmax* (Meehl, 1973d) relies on the following theorem: If three fallible indicator variables are negligibly correlated within a diagnostic taxon and within the extra taxon population, then the covariance of any pair of these is maximized in that class interval on the third indicator that contains the hitmax (optimal, fewest misses) cut on the third indicator. That is, cov(yz) has its largest value for the subset of patients falling in the hitmax interval on x. Starting from this relation we go through a sequence of calculations yielding estimates of the base rate P of the taxon, the frequency distributions of all three of the fallible indicators, the location of all three hitmax cuts, and the inverse probability of taxon membership (via Bayes' theorem) for a patient who has any given combination of the three signs plus or minus.

Our Monte Carlo runs and our single application to a real case in which we know the true answer and pretend not to know it, namely, biological sex diagnosed by three MMPI femininity keys, have been most encouraging and suggest that the method is powerful and quite robust under departures from the simplifying hypotheses. But applying it to a situation in which we do not know the true answer (such as "What is the proportion of unrecognized schizotypes in a mixed psychiatric population?"), how much faith should we have in our numerical results? The best way I know to go about this, since mere replication of the inferred parameter estimates does not answer the question, is by the use of consistency tests. For example, one of the consistency tests in this kind of two-category taxonic situation is this: If we form the product of the differences between the inferred latent means on y and z (schizotypes minus nonschizotypes) and then multiply this product $\Delta\bar{y}\ \Delta\bar{z}$ by the product of the inferred schizotypal base-rate P and its complement Q, then it can be shown that this theoretically calculated quantity should equal the grand covariance of y and z computed directly from the observations. We call this the "total covariance consistency test."

Of course, such a relation is not required to be literally true, because it is known in advance that (a) the impoverished theory has imperfect verisimilitude and (b) all statistical estimates are subject to both systematic and random error. (We are *not* going to do a significance test!) What we have is a problem of robustness and detection of excessive departures from the postulated latent conditions. Golden and I arbitrarily said that we would consider a particular sample as delivering sufficiently accurate information if the estimates of base rate and hit rate were within .10 of the true values, and estimated latent means and standard deviations within one class interval of the truth. (Actually we did much better than that on the average. For example, with sample sizes greater than 400, equal variances, two sigma differences of latent means, and zero intrataxon correlations, the average error for P was only .01 and for latent means and sigmas, less than one fourth standard deviation which is one-half the smallest integral class interval.) But if these tolerances strike you as excessively large, I remind you how much more powerful such numerical claims are in soft psychology than the usual flabby "the boys are taller than the girls" or "the schizophrenics are shyer than the manic

Table 1
Description of Sample Sets

Set	Variable	N	P	M_e	M_t	SD_e	SD_t	D'	SD_t/ SD_e	r	F^a
1.1	N	1,000	.5	8	12	2	2	2	1	0^b	0
1.2		800	.5	8	12	2	2	2	1	0^b	0
1.3		600	.5	8	12	2	2	2	1	0^b	0
1.4		400	.5	8	12	2	2	2	1	0^b	0
2.1	P	1,000	.6	8	12	2	2	2	1	0^b	3
2.2		1,000	.7	8	12	2	2	2	1	0^b	2
2.3		1,000	.8	8	12	2	2	2	1	0^b	8
2.4		1,000	.9	8	12	2	2	2	1	0	0
3.1	D'	1,000	.5	9	12	2	2	1.5	1	0^b	0
3.2		1,000	.5	10	12	2	2	1	1	0^b	15
3.3		1,000	.5	11	12	2	2	.5	1	0	0
3.4		1,000	.5	12	12	2	2	0	1	0	0
4.1	D_t/ SD_e	1,000	.5	8	12	1.9	2.1	2	1.1	0^b	0
4.2		1,000	.5	8	12	1.7	2.3	2	1.3	0^b	0
4.3		1,000	.5	8	12	1.5	2.5	2	1.7	0^b	0
4.4		1,000	.5	8	12	1	3	2	3	0	0
5.1	r	1,000	.5	8	12	2	2	2	1	$.1^b$	0
5.2		1,000	.5	8	12	2	2	2	1	$.3^b$	0
5.3		1,000	.5	8	12	2	2	2	1	$.5^b$	8
5.4		1,000	.5	8	12	2	2	2	1	.8	0
										r_e/r_t	
6.1	$r_e/r_t = 4$	1,000	.8	8	12	2	2	2	1	.5/.125	0
6.2	N	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

Note. N = sample size; P = base rate of the taxon; M_e = mean of the extra taxon class on each indicator; M_t = mean of the taxon on each indicator; SD_e = standard deviation of the extra taxon class on each indicator; SD_t = standard deviation of the taxon on each indicator; $D' = (M_t - M_e)/$ S, where S = $SD_e + SD_t)/2$; r = latent correlation between indicator pairs; F = number of failures of consistency tests in 25 samples.

[a] 94% correct.

[b] Parameter estimates are always or nearly always accurate.

depressives." We then imposed tolerances on each of the four most promising consistency tests derived within the formalism, For example, if the total covariance consistency test T_1 = cov(yz) − PQ $(\bar{y}_s - \bar{y}_n)$ $(\bar{z}_s - \bar{z}_n)$ yields a discrepancy greater than .64 + .74s^2, a "robustness cut" chosen by a combination of analytical derivation with preliminary Monte Carlo trials, then this particular sample is considered "numerically inconsistent" with Consistency Test T_1. Now if any one of the four consistency tests is, so to speak, rejected by a given sample, this is a red flag warning us that we ought not to have much faith in the parametric estimates of interest.

The important question then is, how sensitive are the consistency tests to sample departures from the parametric truth in excess of the tolerance allowed? How often will we draw a sample in which the inferred parameters are in error by more than the tolerance limit imposed but all four consistency tests are satisfied within *their* tolerance limits, leading us mistakenly to trust our results? Second, how often is at least one of the four consistency tests numerically inconsistent (i.e.,

Table 2
Consistency Test Result

	Sample		
Actual situation	Trust-worthy	Sus-picious	Total
Accurate	336	36	372
Inaccurate	0	228	228
Total	336	264	600

outside its tolerance limit) leading us to *mis*trust the sample when in fact all of the sample estimates of the parameters are within their tolerances? The first of these we might call a "false negative" failure on the part of the consistency tests to function jointly; the second is then a false positive.

I restrict my data presentation to Monte Carlo runs in which the samples are generated from a multivariate normal model, although I want to emphasize that our methods are not generally confined to the normal case. Normality was imposed because of Monte Carlo generating problems. In Table 1, the numbers "Set 1.1, 1.2,…" in the first column merely name conditions of fixed population properties and sample sizes, and 25 Monte Carlo samples were drawn per set. The column heads indicate the various population properties, such as taxon base-rate *P*, the two latent taxon means and standard deviations, the mean difference in standard deviation units, the ratio of latent standard deviations, and the within-group correlations. The important result (*F*) indicates how many of the 25 samples under the given set conditions were failures of the consistency tests. Thus, the four consistency tests were applied to each sample, which was classified as probably trustworthy (or probably not) in accordance with the tolerance rules for consistency tests. Then the sample was classified as to whether it was *in fact* trustworthy, that is, whether the main latent parameters were all estimated within their allowed tolerance.

Despite the high average accuracy of our taxometric method when evaluated as mean percent errors in estimating each of the latent parameters (base rate, hit rates, means, standard deviations), if a naive trusting taxometrist relied blindly on the method, hoping to be accurate on all seven parameters on any sample drawn, he would be misled distressingly often were he to lack consistency tests. Among our 600 Monte Carlo samples, all seven latent parameters of the artificial population were estimated to an accuracy within the tolerance levels in 372 samples; that is, on 228 samples at least one parameter was inaccurate. This shows that a trustworthy device for detecting such bad samples is much to be desired. It will not do a taxonomic scientist much good to be "usually quite accurate" if the procedure relied on is nevertheless often (38% of the time) somewhat inaccurate *and the investigator is without a method that warns him when the untoward event has, on a given occasion, occurred.*

In Table 2 the 600 Monte Carlo samples are tallied with respect to each sample's parameter estimation accuracy and whether it passed all four consistency tests. It is encouraging that overall the consistency tests were 94% accurate. Furthermore, the 6% of the samples in which the consistency tests erred were all samples in which they erred conservatively; that is, one or more of the consistency tests was suspiciously outside its tolerance limits, yet none of the latent parameters estimated by the methods was outside *its* tolerance limits. We have not as yet drawn a single Monte Carlo sample (among 600) in which the four consistency tests were conjunctively reassuring but the sample was in fact misleading. This finding suggests that we were unduly stringent, so that if some small amount of leeway were permitted for errors of the other kind, the consistency tests could be somewhat relaxed and, perhaps concurrently, the tolerance limits on the parameter estimates could be somewhat tightened.

There is some interchangeability between original estimators and consistency tests, and the maxcov-hitmax method itself was originally derived by me as a consistency test before I realized that it could better be used as an original search device (see Meehl, Note 3, pp. 28–29; Note 4, pp. 2–6).

Not in reliance on these results, which I present merely as exemplars of a general methodological thesis, I want now to state as strongly as I can a prescription that we should adopt in soft psychology to help get away from the feeble practice of significance test mg: *Wherever possible, two or more nonredundant estimates of the same theoretical quantity should be made, because multiple approximations to a theoretical number are always more valuable, provided that methods of setting permissible tolerances exist, than a so-called exact test of significance, or even an exact setting of confidence intervals*. This is a special case of what my philosopher colleague Herbert Feigl refers to as "triangulation in logical space." It is, as you know, standard procedure in the developed sciences. We have, for instance, something like a dozen independent ways of estimating Avogadro's number, and since they all come out "reasonably close" (again, I have never seen a physicist do a *t* test on such a thing!), we are confident that we know how many molecules there are in a mole of chlorine.

This last point may lead you to ask, "If consistency tests are as important as Meehl makes them out to be, why we don't hear about them in chemistry and physics?" I have a perfect answer to that query. It goes like this: *Consistency tests are so much a part of standard scientific method in the developed disciplines, taken so much for granted by everybody who researches in chemistry or physics or astronomy or molecular biology or genetics, that these scientists do not even bother having a special name for them!* It shows the sad state of soft psychology when a fellow like me has to cook up a special metatheory expression to call attention to something that in respectable science is taken as a matter of course.

Having presented what seems to me some encouraging data, I must nevertheless close with a melancholy reflection. The possibility of deriving consistency tests in the taxonic situation rests on the substantive problems presented by fields like medicine and behavior genetics, and it is not obvious how we would go about doing this in soft areas that are non-taxonic. It may be that the nature of the subject matter in most of personology and social psychology is inherently incapable of permitting theories with sufficient conceptual power (especially mathematical development) to yield the kinds of strong refuters expected by Popperians, Bayesians, and unphilosophical scientists in developed fields like chemistry. This might mean that we could most profitably confine ourselves to low-order inductions, a (to me, depressing) conjecture that is somewhat corroborated by the fact that the two most powerful forms of clinical psychology are atheoretical psychometrics of prediction on the one hand and behavior modification on the other. Neither of these approaches has the kind of conceptual richness that attracts the theory-oriented mind, but I think we ought to acknowledge the possibility that there is never going to be a really impressive theory in personality or social psychology. I dislike to think that, but it might just be true.

Addendum

My colleague, Thomas B. Bouchard, Jr., on reading a draft of this article faulted me for what he saw as a major inconsistency between my neo-Popperian emphasis on falsifiability and my positive assessment of Freud. There is no denying that for such a quantitatively oriented product of the "dust-bowl empiricist" tradition as myself, I do have a soft spot in my heart (Minnesota colleagues would probably say in my head) for psychoanalysis. So, the most honest and straightforward way to deal with Bouchard's complaint might be simply to admit that the evidence on Freud is inadequate and that Bouchard and I are simply betting on different horses. But I can not resist the impulse to say just a bit more on this vexatious question, because while I am acutely aware of a pronounced (and possibly irrational) difference in the "educated prior" I put on Freud as contrasted with rubber band theory or labeling theory or whatever, I am not persuaded that my position is as grossly incoherent as it admittedly appears. Passing the question whether attempts to study psychoanalytic theory by the methods of experimental or differential psychology have on the whole tended to support rather than refute it (see, e.g., Fisher & Greenberg, 1977; Rapaport, 1959; Sears, 1943; Silverman, 1976), my own view is that the best place to study psychoanalysis is the psychoanalytic session itself, as I have elsewhere argued in a far too condensed way (Meehl, 1970/1973e).

I believe that some aspects of psychoanalytic theory are not presently researchable because the intermediate technology required—which really means instruments-cum-theory—does not exist. I mean auxiliaries and methods such as a souped-up, highly developed science of psycholinguistics, and the kind of mathematics that is needed to conduct a rigorous but clinically sensitive and psychoanalytically realistic job of theme tracing in the analytic protocol. This may strike some as a kind of cop-out, but I remind you that Lakatos, Kuhn, Feyerabend, and others have convincingly made the point that there are theories in the physical and biological sciences that are untestable when first propounded because the theoretical and technological development necessary for making certain kinds of observations bearing on them had not taken place. It is vulgar positivism (still held by many psychologists) to insist that any respectable empirical theory must be testable, if testable means *definitively testable right now*.

But I do think that there is another class of consequences of psychoanalytic theory, close to the original 'clinical connections" alleged by Freud, Ferenczi, Jones, Abraham, and others that does not involve much of what Freud called *the witch metapsychology*, where no complicated statistics are needed, let alone the invention of any new formal modes of protocol analysis. Here the problem is mainly that *none of us has bothered to carry out some relatively simple-minded kinds of analyses on a random sample of psychoanalytic protocols collected from essentially naive patients to whom no interpretations have as yet been offered*. This second category is, in my view, a category of research studies that we could have done, but have not done. *Example*: We can easily ascertain whether manifest dream content of a certain kind is statistically associated (in the simple straightforward sense of a patterned fourfold table) with such and such kinds of thematic material in the patient's subsequent associations to the dream. I would not even object to doing significance tests on a batch of such tables, but to explain why would unduly enlarge what is already an addendum.

I cheerfully admit, in this matter, to the presence of at large distance between my subjective personalistic probability (based on my experiences as analysand and practitioner of psychoanalytic therapy) and the present state of the "intersubjective public evidence." That is what I mean by saying that Bouchard and I are betting on different horses. But one must distinguish, as I know from subsequent conversations that he does, between a criticism (a) that what *is* proper evidence *does* presently exist and is *adverse* to a conjecture and (b) an anti-Popperian claim that falsifiability in principle does not matter. If I thought (as does Popper) that Freudian theory was in principle not falsifiable, then I would have to confess to a major inconsistency. But I do think it is falsifiable, although I agree that *some parts* of it cannot *at present* be tested because of the primitive development of the auxiliary theories and the measurement technologies that would be jointly necessary.

A final point on this subject is one that I hesitate to include because it is very difficult to explain in the present state of philosophy of science, and I could be doing my main thesis damage by presenting a cursory and somewhat dogmatic statement of it. Nevertheless, having made the above statements about psychoanalytic theory and having contrasted it favorably with some of the (to me, trivial and flabby) theories in soft psychology, I fear I have an obligation to say it, however ineptly. Once one sees that it is inappropriate to conflate the concepts *rational* and *statistical*, then it is a fuzzy open question, in the present state of the metatheoretician's art, just when a mass of nonquantitative converging evidence can be said to have made a stronger case for a conjecture than the weak kinds of nonconverging quantitative evidence usually represented by the significance testing tradition. I say "when" rather than "whether," because it is blindingly obvious that *sometimes* qualitative evidence of certain sorts is superior in its empirical weight to what a typical social, personality, or clinical psychologist gets in support of a substantive theory by the mere refutation of the null hypothesis. Take, for instance, the evidence in a well-constructed criminal case, such as the evidence that Bruno Hauptmann was the kidnapper of the Lindbergh baby. I do not see how anybody who reads the trial transcript of the Hauptmann case could have a reasonable doubt that he was guilty as charged. Yet I cannot recall any of the mass of data that convicted him as being of a quantitative sort (one cannot fairly except the serial numbers on the gold notes, they being not "measures" but "football numbers").

All of us believe a lot of things that we would not have the vaguest idea how to express as a probability value (*pace* strong Bayesians!) or how to compute as an indirect test of statistical significance. I believe, for instance, that Adolf Hitler was a schizotype; I do not believe that Kaspar Hauser was the son of a prince; I believe that the domestic cat probably was evolved from *Felis lybica* by the ancient Egyptians; I hold that my sainted namesake wrote the letter to the Corinthians but did not write the letter to the Hebrews; I am confident that my wife is faithful to me; and so forth. The point is really a simple one —that there are many areas of both practical and theoretical inference in which nobody knows how to calculate a numerical probability value, and nobody knows how to state the manner or degree in which various lines of evidence converge on a certain conjecture as having high verisimilitude. There are propositions in history (such as, "Julius Caesar crossed the Rubicon") that we all agree are well corroborated by the available documents but without any *t* tests *or the possibility of calculating any*, whereas Fisbee's theory of social behavior is only weakly corroborated by the fact that he got a significant *t* test when he compared the boys and the girls or the older kids and the younger kids on the Hockheimer-Sedlitz Communication Scale. Now I consider my betting on the horse of psychoanalysis to be in the same kind of ball park as my beliefs about Julius Caesar or the evolution of the cat. But, I repeat, this may be a terribly irrational leap of faith on my part. For the purposes of the present article and Bouchard's criticism of it, I hope it is sufficient to say that one could arguably hold that significance testing in soft psychology is a pretentious endeavor that falls under a tolerant neo-Popperian criticism, and could nevertheless enter his personalistic prediction that *when adequate tests become available to us, a sizable portion of psychoanalytic theory will escape refutation*. So I do not think I am actually contradicting myself, but I am personalistically betting on the outcome of a future horse race.

Reference Notes

1. Golden, R., & Meehl, P. E. *Detecting latent clinical taxa, IV: Empirical study of the maximum Covariance method and the normal minimum chi square method, using three MMPI keys to identify the sexes* (Tech. Rep. PR-73-2). Minneapolis: University of Minnesota Psychiatry Research Laboratory, 1973. (a)
2. Golden, R., & Meehl, P. E. *Detecting latent clinical taxa, V: A Monte Carlo study of the maximum covariance method and associated consistency tests* (Tech. Rep. PR-73-3). Minneapolis: University of Minnesota Psychiatry Research Laboratory, 1973. (b)
3. Meehl, P. E. *Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion* (Tech. Rep. PR-65-2). Minneapolis: University of Minnesota Psychiatry Research Laboratory, 1965.
4. Meehl, P. E. *Detecting latent clinical taxa ,II: A simplified procedure, some additional hitmax cut locators, a single-indicator method, and miscellaneous theorems* (Tech. Rep. PR-65-4). Minneapolis: University of Minnesota Psychiatry Research Laboratory, 1968.

References

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

Andreski, S. *Social sciences as sorcery*. London: Deutsch, 1972.

Barker, R. G. *Ecological psychology*. Stanford, Calif.: Stanford Univer. Press, 1968.

Bartlett, M. S. *An introduction to stochastic processes*. Cambridge, Eng.: Cambridge Univer. Press, 1955.

Bolles, R. C. The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, 1962, *11*, 639-645.

Böök, J. A. Genetical aspects of schizophrenic psychoses. In D. D. Jackson (Ed.), *The etiology of schizophrenia*. New York: Basic Books, 1960.

Braithwaite, R. B. *Scientific explanation*. New York: Harper, 1960.

Braun, J. R. (Ed.). *Clinical psychology in transition* (Rev. ed.). Cleveland, Ohio: World, 1966.

Brunswik, E. Representative design and probabilistic theory. *Psychological Review*, 1955, *62*, 236-242.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, *56*, 81-105.

Campbell, N. R. *Foundations of science*. New York: Dover, 1957. (Originally published as *Physics, the elements*, 1920.)

Carnap, R. *Testability and meaning*. New Haven, Conn.: Yale Univer. Graduate Philosophy Club, 1950. (Originally published, l936-1937.)

Carnap, R. Meaning postulates. In R. Carnap, *Meaning and necessity*. Chicago: Univer. of Chicago Press, 1956. (Originally published, 1952.)

Carnap, R. *Philosophical foundations of physics*. New York: Basic Books, 1966.

Cartwright, D. Determinants of scientific progress: The case of the risky shift. *American Psychologist*, 1973, *28*, 222-231.

Cronbach, L. J. The two disciplines of scientific psychology. *American Psychologist*, 1957, *12*, 671-684.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: Univer. of Minnesota Press, 1973. (Originally published, 1955.)

Eisberg, R. M. *Fundamentals of modern physics*. New York: Wiley, 1961.

Feigl, H. Some major issues and developments in the philosophy of science of logical empiricism, In H. Feigl & M. Scriven (Eds,), *The foundations of science and the concepts of psychology and psychoanalysis: Minnesota studies in the philosophy of science* (Vol. 1). Minneapolis: Univer. of Minnesota Press, 1956.

Feller, W. *An introduction to probability theory and its applications* (2nd ed.). New York: Wiley, 1957.

Feyerabend, P. K. Explanation, reduction, and empiricism. In H. Feigl & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science* (Vol. 3): *Scientific explanation, space and time*. Minneapolis: Univer. of Minnesota Press, 1962.

Feyerabend, P. K. Problems of Empiricism, Part I. In R. G. Colodny (Ed.), *Beyond the edge of certainty*. Englewood Cliffs, N.J.: Prentice-Hall, 1965.

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

Feyerabend, P. K. Problems of empiricism, Part II. In R. G. Colodny (Ed.), *The nature and function of scientific theories*. Pittsburgh, Pa.: Univer. of Pittsburgh Press, 1971.

Fisher, F. M. *The identification problem in econometrics*. New York: McGraw-Hill, 1966.

Fisher, R. A. *Statistical methods and scientific inference*. Edinburgh, Scotland: Oliver & Boyd, 1956.

Fisher, R. A. *The design of experiments* (8th ed.). Edinburgh, Scotland: Oliver & Boyd, 1966.

Fisher, R. A. *Statistical methods for research workers* (13th ed.), Edinburgh, Scotland: Oliver & Boyd, 1967.

Fisher, S., & Greenberg, P. *The scientific credibility of Freud's theories and therapy*. New York: Basic Books, 1977.

Fiske, D. W. The limits of the conventional science of personality. *Journal of Personality*, 1974, *24*, 1-11.

Gergen, K. J. Social psychology as history. *Journal of Personality and Social Psychology*, 1973, *26*, 309-320.

Golden, R. *Psychometric verisimilitude*. Unpublished doctoral dissertation, Univer. of Minnesota, 1976.

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

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

Gottesman, I. I. Heritability of personality: A demonstration. *Psychological Monographs*, 1963, *77*(9, Whole No. 572).

Gottesman, I. I., & Shields, J. *Schizophrenia and genetics: A twin study vantage point*. New York: Academic Press, 1972.

Grünbaum, A. The Duhemian argument. *Philosophy of Science*, 1960, *11*, 75-87.

Grünbaum, A. Falsifiability of theories: Total or partial? *Synthese*, 1962, *14*, l7-34.

Grünbaum, A. Can we ascertain the falsity of a scientific hypothesis? *Stadium Generale*, 1969, *22*, 1061-1093.

Grünbaum, A. Ad hoc auxiliary hypotheses and falsificationism. *British Journal for the Philosophy of Science*, 1976, *27*, 329-362.

Gunther, M. *The luck factor*. New York: Macmillan, 1977.

Harrison, T. R., et al. (Eds.), *Principles of internal medicine*. New York: McGraw-Hill, 1966.

Hempel, C. G. *Fundamentals of concept formation in empirical science*. Chicago: Univer. of Chicago Press, 1952.

Hempel, C. G. The theoretician's dilemma. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science* (Vol. 2): *Concepts, theories, and the mind-body problem*. Minneapolis: Univer. of Minnesota Press, 1958.

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

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

Hinde, R. A. *Animal behavior* (2nd ed.). New York: McGraw-Hill, 1970.

Hogan, R., DeSoto, C. B., & Solano, C. Traits, tests, and personality research, *American Psychologist*, 1977, *32*, 255-264.

Holt, R. R. Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology*, 1958, *56*, 1-12.

Holton, G. *Thematic origins of scientific thought: Kepler to Einstein*. Cambridge, Mass.: Harvard Univer. Press, 1973.

Jencks, C. *Inequality*. New York: Basic Books, 1972.

Johnson, W. E. *Logic, Part I*. New York: Dover Publications, 1964. (Originally published, 1921.)

Kemeny, J. G., Snell, J. L., & Thompson, G. L. *Introduction to finite mathematics*. Englewood Cliffs, N.J.: Prentice-Hall, 1957.

Keuth, H. On prior probabilities of rejecting statisticcal hypotheses. *Philosophy of Science*, 1973, *40*, 538-546.

Kordig, R. The comparability of scientific theories. *Philosophy of Science*, 1971, *38*, 467-485.

Kordig, C. R. Discussion: Observational invariance. *Philosophy of Science*, 1973, *40*, 555-569.

Kuhn, T. S. Logic of discovery or psychology of research? In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge, Eng.: Cambridge Univer. Press, 1970. (a)

Kuhn, T. S. Reflections on my critics. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge, England: Cambridge Univer. Press, 1970. (b)

Kuhn, T. S. *The structure of scientific revolutions* (2nd ed.). Chicago: Univer. of Chicago Press, 1970. (c)

Lakatos, I. Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge, England: Cambridge Univer. Press, 1970.

Lakatos, I. Popper on demarcation and induction. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper* (Vol. 1). LaSalle, Ill.: Open Court, 1974. (a)

Lakatos, I. The role of crucial experiments in science. *Studies in History and Philosophy of Science*, 1974, *4*, 309-325. (b)

Langmuir, I. Science, common sense and decency. *Science*, 1943, *97*,1-7.

Lazarus, A. A. *Behavior therapy and beyond*. New York: McGraw-Hill, 1971.

Lewis, D. How to define theoretical terms. *Journal of Philosophy*, 1970, *67*, 427-446.

Li, C. C. *Path analysis*. Pacific Grove, Calif.: Boxwood Press, 1975.

Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports Monograph*, 1957, 9, 635-694.

London, I. D. Some consequences for history and psychology of Langmuir's concept of convergence and divergence of phenomena. *Psychological Review*, 1946, *53*, 170-188.

MacCorquodale, K., & Meehl, P. E. Edward C. Tolman. In W. K. Estes et al., *Modern learning theory: A critical analysis of five examples*. New York: Appleton-Century-Crofts, 1954.

Maxwell, G. Meaning postulates in scientific theories. In H. Feigl & G. Maxwell (Eds.), *Current issues in the philosophy of science*. New York: Holt, Rinehart & Winston, 1961.

Maxwell, G. The necessary and the contingent. In H. Feigl & G. Maxwell (Eds.), *Scientific explanation, space and time: Minnesota studies in the philosophy of science* (Vol. 3). Minneapolis: Univer. of Minnesota Press, 1962.

Maxwell, G. Corroboration without demarcation. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper*. LaSalle, Ill.: Open Court, 1974.

McGuire, W. J. The yin and yang of progress in social psychology: Seven koans. *Journal of Personality and Social Psychology*, 1973, *26*, 446-456.

Meehl, P. E. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: Univer. of Minnesota Press, 1954.

Meehl, P. E. Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science* (Vol. 4) *Analyses of theories and methods of physics and psychology*. Minneapolis: Univer. of Minnesota Press, 1970. (a)

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

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

Meehl, P. E. The cognitive activity of the clinician. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: Univer. of Minnesota Press, 1973. (a) (Originally published, 1960.)

Meehl, P. E. High school yearbooks: A reply to Schwarz. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: Univer. of Minnesota Press, 1973. (b) (Originally published, 1971.)

Meehl, P. E. Schizotaxia, schizotypy, schizophrenia. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: University of Minnesota Press, 1973. (c) (Originally published, 1962.)

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

Meehl, P. E. Some methodological reflections on the difficulties of psychoanalytic research. *Psychological Issues*, 1973, *8*, 104-115. (e) (Originally published, 1970.)

Meehl, P. E. Some ruminations on the validation of clinical procedures. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: Univer. of Minnesota Press, 1973. (f) (Originally published, 1959.)

Meehl, P. E. Specific genetic etiology, psychodynamics, and therapeutic nihilism. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: Univer. of Minnesota Press, 1973. (g) (Originally published, 1972.)

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

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

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

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

Miachel, W. On the future of personality measurement. *American Psychologist*, 1977, *32*, 246-254.

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

Nagel, F. *The structure of science*. New York: Harcourt, Brace & World, 1961.

Nambu, Y. The confinement of quarks. *Scientific American*, 1976, *235*, 48-60.

Nozick, R. *Anarchy, state and utopia*. New York: Basic Books, 1974.

Oakes, W. F. On the alleged falsity of the null hypothesis. *Psychological Record*, 1975, *25*, 265-272.

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

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

Pap, A. *An introduction to the philosophy of science*. New York: Free Press, 1962.

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

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

Popper, K. R. *Objective knowledge*. Oxford, Eng.: Oxford Univer. Press, 1972.

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

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

Ramsey, F. P. Theories. In R. R. Braithwaite (Ed.), *F. P. Ramsey's The foundations of mathematics and other logical essays*. Paterson, N.J.: Littlefield, Adams, 1960. (Orignially published, 1931.)

Rapaport, D. The structure of psychoanalytic theory: A systematizing attempt. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 3). *Formulations of the person and the social context*. New York: McGraw-Hill, 1959.

Read, R. C. *A mathematical background for economists and social scientists*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Schilpp, P. A (Ed.), The phi1osophy of Karl Popper. LaSalle, Ill.: Open Court, 1974.

Schlenker, B. R. Social psychology and science. *Journal of Personality and Social Psychology*, 1974, *29*, 1-15.

Sears, R. R. Survey of objective studies of psychoanalytic concepts. New York: *Social Science Research Council Bulletin* No. 51, 1943.

Sellars, W. Concepts as involving laws and inconceivable without them. *Philosophy of Science*, 1948, *15*, 287-315.

Sells, S. B. An interactionist looks at the environment. *American Psychologist*, 1963, *18*, 696-702.

Silverman, L. H. Psychoanalytic theory: "The reports of my death are greatly exaggerated." *American Psychologist*, 1976, *31*, 621-637.

Skinner, B. F. *The behavior of organisms*. New York: Appleton-Century, 1938.

Slater, E. The monogenic theory of schizophrenia. In J. Shields & I. I. Gottesman (Eds.), *Man, mind, and heredity: Selected papers of Eliot Slater on psychiatry and genetics*. Baltimore, Md.: Johns Hopkins Univer. Press, 1971. (Originally published, 1958.)

Smith, M. B. Criticisms of a social science. *Science*, 1973, *180*, 610-612.

Stoddard, L. *Luck, your silent partner*. New York: Liveright, 1929.

Swoyer, C., & Monson, T. C. Theory confirmation in psychology. *Philosophy of Science*, 1975, *42*, 487-502.

Vico, G. *The new science of Giambattista Vico* (3rd rev. ed.) (T. C. Bergin and M. H. Fisch, trans.). Ithaca: Cornell Univer. Press, 1948. (Originally published, 1744.)

Waismann, F. Verifiability. *Proceedings of the Aristotelian Society*, 1945, *19*, 119-150.

Wiggins, J. Despair and optimism in Minneapolis. *Contemporary Psychology*, 1973, *18*, 605-606.

Received December 12, 1977.

Format version: April 2002
pdf by ljy August 2002