Psychological Reports, 1990, 66, 195–244 © Psychological Reports 1990
Monograph Supplement 1–V66

WHY SUMMARIES OF RESEARCH ON PSYCHOLOGICAL THEORIES ARE OFTEN UNINTERPRETABLE[1, 2]

PAUL E. MEEHL

University of Minnesota

Summary.—Null hypothesis testing of correlational predictions from weak substantive theories in soft psychology is subject to the influence of ten obfuscating factors whose effects are usually (1) sizeable, (2) opposed, (3) variable, and (4) unknown The net epistemic effect of these ten obfuscating influences is that the usual research literature review is well nigh uninterpretable Major changes in graduate education, conduct of research, and editorial policy are proposed

Recently, I read an article in the *Psychological Bulletin* summarizing the research literature on a theory in personology. I had some interest in it both for its intrinsic importance and because the theorist is an old friend and former academic colleague. The reviewer seemed scrupulously fair in dealing with the evidence and arguments, and I do not believe any reader could discern even faint evidence of bias pro or con. The empirical evidence on this theory has now accumulated to a considerable mass of factual reports and associated theoretical inferences, so we are not dealing with a recently advanced conjecture on which the evidence is sparse in amount or confined to too narrow a fact domain. Despite this large mass of data and the scholarly attributes of the reviewer, upon completing the reading I found myself puzzled as to what a rational mind ought to conclude about the state of the evidence. Given all these facts and arguments based upon them, pulled together by a reviewer of competence and objectivity, am I prepared to say that my friend X's theory has been refuted, or strongly corroborated, or is in some vague epistemic region in between? If, taken as it stands, the theory seems to have been refuted, is it nevertheless doing well enough considering the whole fact domain and the plausible explanations of some seeming predictive failures, that we should continue to investigate it and try to patch it up (i.e., does it seem to have enough verisimilitude to warrant occupying psychologists with amending it so its verisimilitude may increase)? Or, is the state of the evidence such a mess conceptually and interpretatively that perhaps the thing to do is to give it up as a bad job and start working on something else?

[1] This paper was originally prepared in 1985 as a contribution to *Improving inquiry in social science A volume in honor of Lee J Cronbach,* edited by R. E. Snow and D. F. Wiley, to be published by Lawrence Erlbaum Associates, Hillsdale, NJ

[2] Send reprint requests to the author, University of Minnesota, Department of Psychology, 75 East River Road, Minneapolis, MN 55455

Inquiry among my colleagues suggests that this befuddled state following the reading of a research literature review is not peculiar to me, or even a minority of faculty in a first-class psychology department, but is so frequent as to be almost the norm. Why is this? This personal phenomenon of cognitive bafflement writ large is, of course, the well known deficiency of most branches of the social sciences to have the kind of cumulative growth and theoretical integration that characterizes the history of the more successful scientific disciplines. I do not here address the question whether psychology and sociology are really in poorer shape as regards replication of findings than chemistry or astronomy, although I am aware that there is a minority report on that score. In what follows I shall presuppose that, by and large, with certain striking exceptions (which I think are rather easy to account for as exceptions), theories in the "soft areas" of psychology have a tendency to go through periods of initial enthusiasm leading to large amounts of empirical investigation with ambiguous over-all results. This period of infatuation is followed by various kinds of amendment and the proliferation of *ad hoc* hypotheses. Finally, in the long run, experimenters lose interest rather than deliberately discard a theory as clearly falsified. As I put it in a previous paper on this subject (1978), theories in the "soft areas" of psychology have a fate like Douglas MacArthur said of what happens to old generals, "They never die, they just slowly fade away." The optimistic reader who does not agree with this assessment may still find the material that follows of interest because much of it bears upon the improvement of research and interpretation.

The discussion that follows, except as specifically noted otherwise, is confined to surveys of research evidence sharing three properties, to wit, (a) theories in so called "soft areas," (b) data correlational, and (c) positive findings consisting of refutation of the null hypothesis. I do not offer a precise specification of "soft area," which is not necessary for what I am doing here, but I am sure the reader knows approximately what branches of psychology are normally so classified. I will content myself with listing the chief ones, namely, clinical, counseling, personality theory, and social psychology. Let me emphasize that I am concerned here wholly with the testing of *explanatory theories* in these areas, and that most of what I say does not apply to purely technological generalizations such as the question whether a certain Rorschach sign is statistically predictive of suicide risk, or that tall mesomorphic males make better military leaders on the average. By 'correlational' I mean simply that the lawful relationship obtained in the observations is based upon calculating a statistic on cross-sectional data, taking the organisms as they come, rather than experimentally manipulating certain factors while other factors are held constant or subjected to a randomizing process. By property (c), I mean that the theory under scrutiny is not powerful enough to generate a numerical point value, from which an experimental finding may or may not deviate significantly, but is so weak that it merely implies that one group will score higher than another, or that there is some nonzero cross-sectional correlation between two measured variables.

It is important to take explicit notice of the methodological point involved in property (c). Strong theories leading to numerical predictions are subjected to danger of falsification by a positive significance test, which is the way the equivalent of “statistical significance” is commonly used in chemistry, physics, astronomy, and genetics. Weak theories only predict a directional difference or an association between two things without specifying its size within a narrow range of values, so that the way in which a significance test is employed by psychologists and sociologists is precisely the reverse from its use in hard science. This leads to the paradox that an enhancement of statistical power, say by improvement of the logical design, increased reliability of the measures, or increased sample size has precisely the opposite effect in soft psychology from the one that it has in physics (Meehl, 1967).

An important extension of condition (b) is experimental research in which, while causal factors are manipulated by the experimenter and subjects assigned to treatments on the basis of some mixture of equated factors and randomization, a crucial feature of the statistical analysis is an interaction effect between the manipulated factor and an attribute (trait, demographic, life-history, psychometric, or whatever) of the individuals. When experimental interpretation hinges upon the presence of such an interaction, so that the main effect induced by the manipulated variable, taken by itself, does not suffice to test the substantive theory, such an experimental study is classified as “correlational” in sense (b) above, and the criticisms below apply with the full force that they have in a purely correlational (nonmanipulative) investigation.

With these rough stipulations, I propound and defend a radical and disturbing methodological thesis. *Thesis: Null hypothesis testing of correlational predictions from weak substantive theories in soft psychology is subject to the influence of ten obfuscating factors whose effects are usually (1) sizeable, (2) opposed, (3) variable, and (4) unknown. The net epistemic effect of these ten obfuscating influences is that the usual research literature review is well-nigh uninterpretable.* I want to emphasize that I am not about to offer a list of nit-picking criticisms of the sort that we used to hear from some statisticians when I was a graduate student as, for example, that somebody used a significance test that presupposes normality when the data were not exactly normal, or the old hassle about one-tail versus two-tail significance testing, or, in the case of experiments having higher order interactions, the argument about whether some of these sums of squares are of such marginal significance (and so uninterpretable theoretically) that they should really be pooled as part of the error term in the denominator. I am making a claim much stronger than that, which is I suppose the main reason that students and colleagues have trouble hearing it, since they might not know what to do next if they took it seriously! The italicized thesis above is stated strongly, but I do not exaggerate for emphasis, I mean it quite literally. I mean that my befuddled response upon reading that literature review of my friend's theory is not due to Meehl being obsessional, senile, or statistically inept, but is precisely the right response of a rational mind, given the combined operation of the ten obfuscating factors that I am about to explain. These obfuscating factors are not typically of negligible size, although in a particular case one or two of them may not be very large, *but we do not know which ones*. They vary from one domain and from one experiment and from one measuring instrument to another, but we do not typically know how big a given one is in a given setting. About half of them operate to make good theories look bad, and the other half tend to make poor theories look good, and at least one of the factors can work either way. Because of these circumstances, I take my thesis above to be literally true. The combined operation of the ten factors — powerful, variable, unmeasured, and working in opposition to each other, counterbalancing one another's influence with an indeterminate net result — makes it impossible to tell what a "box score" of statistical significance tests in the research literature proves about the theory's verisimilitude. If the reader is impelled to object at this point "Well, but for heaven's sake, you are practically saying that the whole tradition of testing substantive theories in soft psychology by null hypothesis refutation is a mistake, despite R. A. Fisher and Co. in agronomy," that complaint does not disturb me because that is *exactly* what I am arguing.

This paper will not treat all philosophy of science aspects of the topic; I confine my remarks here to a brief statement of the usual situation in testing a substantive theory by predicting some observational relationship, which is good enough for present purposes and is not, I think, controversial in the relevant aspects. All logicians and historians of science agree upon the essentials. Theories do not entail particulars, that is, single observations; but theories taken with a statement of conditions entail relations between particulars. The derivation of a prediction about observational facts[3] involves, when spelled out in detail, a conjunction of several premises, and this will I think always be true in the testing of theories in soft psychology. At least I am not aware of any exceptions. Let the substantive theory of interest be *T*. We have

[3] For present purposes I take "observational" to be unproblematic although in strict epistemology it remains a knotty question.

one or more auxiliary theories (some of which may be about instrumentation and others about the psyche) which are not the main focus of the investigator's interest, A_1, A_2, …. Then we need a negative statement which is not formulated with concrete content like an auxiliary but which says "other things being equal." Following Lakatos (1970, 1974), I shall refer to this simply as the *ceteris paribus* clause, designated C_p. Finally, we have statements about the experimental conditions, designated C_n for "conditions," achieved either by manipulation or, in the case of individual differences variables (traits, test scores, demographics, life history facts), by selection of subjects. In other words, for the derivation we must trust that the investigator did what he said he did in getting the subjects and in doing whatever he did to the subjects. Then, if (O_1, O_2) are observational statements, the structural model for testing a substantive theory looks like this:[4]

Derivation of observational conditional: $T \bullet A_1 \bullet A_2 \bullet C_p \bullet C_n \rightarrow (O_1 \supset O_2)$

Theoretical risk: The prior probability $p(O_1|O_2)$, absent theory, should be small (cf. Popper, 1959, 1962, 1983; Schilpp, 1974).

I now present and briefly discuss, without rigorous "hammer blow" proofs in all cases but hopefully with sufficient persuasiveness, the ten obfuscating factors that make H_0-refutation in the soft areas largely uninterpretable:

1. *Loose derivation chain*: Very few derivation chains running from the theoretical premises to the predicted observational relation are deductively tight. Logicians and historians of science have pointed out that this is even true in the "exact" mathematicized sciences such as theoretical physics. *A fortiori* there are few tight, rigorous deductions in most areas of psychology, and almost none in soft psychology. While the theorist or the experimenter may present a tight derivation for certain portions of the prediction (e.g., those that involve a mathematical model), he often relies upon one or more "obvious" inferential steps which, if spelled out, would require some additional unstated premises. These unstated premises are of an intuitive, commonsensical, or clinical experiential nature, and sometimes involve nothing

[4] In what follows, the logician's dot '•' denotes *conjunction* ("and"); the arrow '→' denotes *deductive derivability* ("entails," "causally implies"); the horseshoe '⊃' denotes the *material conditional* ("If …, then …"), without entailment. That is, O_2 does not directly follow from O_1 but the combination ($O_1 \bullet {\sim}O_2$) is impossible if the conjunction to the arrow's left is granted. In the logical notation on p. 200, the tilde ' ~ ' denotes *negation* ("not"); the wedge '∨' denotes disjunction ("either … or …," and maybe both); and the symbol '∴' means "therefore."

more complicated than reliance upon the ordinary semantics of trait names. I do not wish to be understood as criticizing this, but I am listing sources of logical slippage and this is obviously one of them. To the extent that the derivation chain from the theory and its auxiliaries to the predicted factual relation is loose, a falsified prediction cannot constitute a strict, strong, definitive falsifier of the substantive theory. The extent to which a successful prediction mediated by such a loose derivation chain with unstated premises *supports* the theory is somewhat more difficult to assess, and I will pass on this for now.

2. *Problematic auxiliary theories*: Here the auxiliary theories, whether of instrumentation or about the subject matter proper, are not suppressed as unstated premises but explicitly stated. Now in soft psychology it sometimes happens — arguably as often as not — that each auxiliary theory is itself nearly as problematic as the main theory we are testing. When there are several such problematic auxiliary theories, the joint probability that they all obtain may be considerably lower than the prior probability of the substantive theory of interest. Here again, the valid form of the syllogism involved in a refutation reads:

$$T \bullet A_1 \bullet A_2 \bullet C_p \bullet C_n \rightarrow\bullet (O_1 \supset O_2)$$

$$(O_1 \bullet {\sim}O_2) \qquad \text{Observational result.}$$

$$\therefore\ {\sim}(T \bullet A_1 \bullet A_2 \bullet C_p \bullet C_n), \qquad \text{Formally equivalent to}$$

$${\sim}T \vee {\sim}A_1 \vee {\sim}A_2 \vee {\sim}C_p \vee {\sim}C_n$$

so that what we intended to refute if the predictions didn't pan out was T, but the logical structure leaves us not knowing whether the prediction failed because T was false or because one or more of the conjoined statements A_1, A_2, C_p, or C_n were false. This reasoning applies as well to Sections 3 and 4 following.

3. *Problematic* ceteris paribus *clause*: The *ceteris paribus* clause does not, of course, mean by "everything else being equal . . ." that the individual subjects are all equated; in a typical study in soft psychology they are definitely not and the individual differences among them appear in the denominator of the significance test. What is meant here by *ceteris paribus* when we test a theory by showing a statistical relationship to be nonzero (significant t, F, r, χ^2), is that while the individuals vary in respect to those factors that we have not controlled but allowed to vary (and which we hope are, therefore, "taken care of" in the statistics by randomization), the alleged causal influence does not in some significant subset of subjects have an additional effect operating systematically in a direction opposed to the one that our theory and the auxiliaries intend. *Example*: Suppose I am trying to test a theory about the difference between introverts and extraverts in their need

for social objects in relationship to experienced stress. I am manipulating a factor experimentally by an operation intended to induce anxiety. But this "experimental study" falls under the present thesis because I am examining the substantive theory via the interaction between the experimental factor and an individual differences variable that the subjects bring with them, to wit, their scores on a social introversion questionnaire. To make the subjects anxious without physical pain I give the experimental group misinformation that they got a grade of "F" on the midquarter exam. I administer a projective test, say the TAT, scored for current regnant *n Affiliation*. The critical (statistical) analysis bearing on the theory is the interaction between social introversion as measured by the questionnaire and the effect of the experimental stress on the output variable, TAT affiliative score. Every psychologist realizes immediately that there is a problematic auxiliary theory involved here, namely, the psychometric validity of the TAT as a projective measure of currently regnant affiliative motives. Further, there is the auxiliary theory that telling people they failed a midquarter will make them anxious. Surely nobody will deny that these are both highly problematic, arguably as problematic as the initial theory itself.

As regards the *ceteris paribus* clause, it may be that telling an "A" student that he failed the midquarter will result either in his disbelief or in some cases in a response of resentment, since he knows he did better and therefore somebody must have done a bum job of scoring or made a clerical mistake. What will such induced anger do to the kinds of TAT stories he tells about human subjects? On the other hand, a poor student may not have his anxiety mobilized very much by the reported "F," since for him that's par for the course.

It is clear that we have a couple of highly problematic auxiliaries, one on the input and one on the output side, together with a problematic *ceteris paribus* clause. If the impact of the grade misinformation is a strong correlate of an individual differences variable that's *not explicitly part of the design*, there may be a subset of individuals for whom the scored TAT behavior is suppressed by the induced state of rage, even if the misinformation has also produced in most or all of those subjects an increase in anxiety.

I don't think any psychologist would consider these unpleasant possibilities the least bit far-fetched. The main point is that the joint problematicity of (1) the auxiliary theories about the psychometric validity of the TAT, (2) the adequacy of the grade misinformation as an eliciter of strong anxiety in all of the subjects, and (3) the *ceteris paribus* that there are no significant correlations between individual differences variables of the subjects (including perhaps their social introversion as well as their usual scholastic performance) and the affective or cognitive states induced, could to some extent countervail the induced state on which we have our eye.

4. *Experimenter error*: Here I refer on the input side to an imperfect realization of the particulars, experimenter mistakes in manipulation. Perhaps the investigator's research assistant is enthusiastic about her theory and without consciously intending — I am completely omitting conscious faking of data — slants the way the grade information is given or picks up cues about anxiety while administering the TAT and consequently shuts off the stories a little quicker or whatever. Exactly how much experimenter error occurs either in experimental manipulation or experimenter bias in recording observations is still in dispute, but again no knowledgeable psychologist would say that it is so rare as to be of zero importance (Rosenthal, 1966; see also Mahoney, 1976).

5. *Inadequate statistical power*: It is remarkable, and says something discouraging about the sociology of science, that more than a quarter century after Jacob Cohen's classic paper on the power function in abnormal-social psychology research (1962), only a minority of investigators mention the statistical power function in discussing experimental design or interpreting data, although naturally there is some temptation to allude to it when explaining away those failures to reach statistical significance whenever facing a mixed bag of results.[5] Because of its special role in the social sciences and the fact that like some of the other obfuscators it does lend itself to some degree of quantitative treatment, I have listed inadequate statistical power separately. A philosopher of science might justifiably argue that statistical power should be included in the guise of an auxiliary.

Because I believe that one of the commonest confusions in thinking about statistical significance testing lies in the conflation of substantive and statistical hypotheses, I throughout use the word "theory" or the phrase "substantive theory" to designate the conjectured processes or entities being studied, and "hypothesis" to refer to the statistical hypothesis that allegedly flows from this substantive picture. I do not know quite how to go about formulating the statistical power function as an auxiliary, and for that reason and its special role in social science I list it separately. Despite Cohen's empirical summary and his tentative recommendations, one has the impression that many psychologists and sociologists view this whole line of concern as a kind of nit-picking statistician's refinement, or perhaps a "piece of friendly advice to researchers," which they may or may not elect to act upon. A moment's reflection tells us that this last view is a grave methodological mistake. Suppose that a substantive theory has perfect verisimilitude, ditto the auxiliaries and the *ceteris paribus* clause; but, as is usual in soft psychology, the substantive theory does not make a numerical point prediction as to the *size* of, say, a correlation coefficient between two observational measures.

[5] I have not done a formal count, but leafing through a few issues of any journal in the soft areas will convince the reader that it is a *very small* minority.

However, suppose that, if pressed, the theorist will tell us that if his theory has high verisimilitude a correlation of at least .35 is expectable, and that he would consider a lower value as hardly consistent with his causal model. Then any correlation larger than .35 will count as a corroborator of the theory, and, to play the scientific game fairly (whether we are Popperians or not, we have given Popper this much about *modus tollens*!), a failure to find a correlation $r > .35$ constitutes a refutation. That is, the theorist says that, while he doesn't know just how big the correlation is, it ought to be at least that big; otherwise he would admit that the facts, if not totally slaying the theory immediately, speak strongly against it. At the lower end of this region of allowable true values, suppose our sample size is such that we would have only a 60% statistical power at the conventional .05 significance level. Surely this is not some minor piddling defect in the study, it is a gross abuse of the theory taken as it stands.

Imagine a chemist who, relying on the old fashioned litmus test for whether something is an acid, told us that the test papers he uses are unfortunately only about 60% blue litmus (that one expects to turn red) and the other 40% are phenolphthalein papers (which, if I recall my undergraduate chemistry, turn red in the presence of a base). If it were important for us to know whether something was acidic or basic, or simply whether it was an acid or not an acid, how dependable would we think the work of a chemist who drew test slips from a jar, 40% of which he can foresee will give the wrong answer, so that of 100 batches of substances studied (even if they were all acid as predicted by some theory) we would get a box score of only 60 to 40? In chemistry such an approach would be considered scandalous.

It will not do to say that such an approach is excusable in psychology because chemistry is easier to do than psychology. Unlike some of the other obfuscators in my list, the establishment of sufficient statistical power can easily be achieved, as Cohen pointed out, so that deficient power is not a plausible explanation of a falsifying result. I am really at a loss to understand the sociology of this matter in my profession. Of course if people never claimed that they had proven the null hypothesis by failing to refute it (taking some of Fisher's injunctions quite literally), this would not be so serious. As we know, Fisher himself was not very tractable on the subject of power, although he got at it in his discussion of precision. But despite the mathematical truth of Fisher's point (which we can sidestep by changing from a point to a range hypothesis and formulating the directional null hypothesis that way), a null result, a failure to reach significance, is regularly counted against a theory. Despite Fisher, it is hard to see how psychologists could do otherwise: if we only count the pluses and ignore the minuses, it is a foregone conclusion that all theories will be corroborated, including those that have no verisimilitude at all.

6. *Crud factor*: In the social sciences and arguably in the biological sciences, “everything correlates to some extent with everything else.” This truism, which I have found no competent psychologist disputes given five minutes reflection, does not apply to pure experimental studies in which attributes that the subjects bring with them are not the subject of study (except in so far as they appear as a source of error and hence in the denominator of a significance test).[6] There is nothing mysterious about the fact that in psychology and sociology everything correlates with everything. Any measured trait or attribute is some function of a list of partly known and mostly unknown causal factors in the genes and life history of the individual, and both genetic and environmental factors are known from tons of empirical research to be themselves correlated. To take an extreme case, suppose we construe the null hypothesis literally (objecting that we mean by it “almost null” gets ahead of the story, and destroys the rigor of the Fisherian mathematics!) and ask whether we expect males and females in Minnesota to be precisely equal in some arbitrary trait that has individual differences, say, color naming. In the case of color naming we could think of some obvious differences right off, but even if we didn’t know about them, what is the causal situation? If we write a causal equation (which is not the same as a regression equation for pure predictive purposes but which, if we had it, would serve better than the latter) so that the score of an individual male is some function (presumably nonlinear if we knew enough about it but heresupposed linear for simplicity) of a rather long set of causal variables of genetic and environmental type X_1, X_2, … X_m. These values are operated upon by regression coefficients b_1, b_2, …b_m.

Now we write a similar equation for the class of females. Can anyone suppose that the beta coefficients for the two sexes will be exactly the same? Can anyone imagine that the mean values of all of the *X*s will be exactly the same for males and females, even if the culture were not still considerably sexist in child-rearing practices and the like? If the betas are not exactly the same for the two sexes, and the mean values of the *X*s are not exactly the same, what kind of Leibnitzian preestablished harmony would we have to imagine in order for the mean color-naming score to come out exactly equal between males and females? It boggles the mind; it simply would never happen. As Einstein said, “the Lord God is subtle, but He is not malicious.” We cannot imagine that nature is out to fool us by this kind of delicate balancing. Anybody familiar with large scale research data takes it as a matter of course that when the *N* gets big enough she will not be looking for the

[6] My colleague, David Lykken, and several high-caliber graduate students who have heard me lecture on this topic hold that I am too conservative in confining my “obfuscator thesis” to correlational research, and they make a strong if not to me persuasive case, but I set that aside with this mere mention of it in the present context.

statistically significant correlations but rather looking at their patterns, since almost all of them will be significant. In saying this, I am not going counter to what is stated by mathematical statisticians or psychologists with statistical expertise. For example, the standard psychologist's textbook, the excellent treatment by Hays (1973, page 415), explicitly states that, taken literally, the null hypothesis is always false.

Twenty years ago David Lykken and I conducted an exploratory study of the crud factor which we never published but I shall summarize it briefly here. (I offer it not as "empirical proof" — that H_0 taken literally is quasi-always false hardly needs proof and is generally admitted — but as a punchy and somewhat amusing example of an insufficiently appreciated truth about soft correlational psychology.) In 1966, the University of Minnesota Student Counseling Bureau's Statewide Testing Program administered a questionnaire to 57,000 high school seniors, the items dealing with family facts, attitudes toward school, vocational and educational plans, leisure time activities, school organizations, etc. We cross-tabulated a total of 15 (and then 45) variables including the following (the number of categories for each variable given in parentheses): father's occupation (7), father's education (9), mother's education (9), number of siblings (10), birth order (only, oldest, youngest, neither), educational plans after high school (3), family attitudes towards college (3), do you like school (3), sex (2), college choice (7), occupational plan in ten years (20), and religious preference (20). In addition, there were 22 "leisure time activities" such as "acting," "model building," "cooking," etc., which could be treated either as a single 22-category variable or as 22 dichotomous variables. There were also 10 "high school organizations" such as "school subject clubs," "farm youth groups," "political clubs," etc., which also could be treated either as a single ten-category variable or as ten dichotomous variables. Considering the latter two variables as multichotomies gives a total of 15 variables producing 105 different cross-tabulations. All values of χ^2 for these 105 cross-tabulations were statistically significant, and 101 (96%) of them were significant with a probability of less than 10^{-6}.

If "leisure activity" and "high school organizations" are considered as separate dichotomies, this gives a total of 45 variables and 990 different cross-tabulations. Of these, 92% were statistically significant and more than 78% were significant with a probability less than 10^{-6}. Looked at in another way, the median number of significant relationships between a given variable and all the others was 41 out of a possible 44!

We also computed MCAT scores by category for the following variables: number of siblings, birth order, sex, occupational plan, and religious preference. Highly significant deviations from chance allocation over categories were found for each of these variables. For example, the females score higher than the males; MCAT score steadily and markedly decreases with increasing numbers of siblings; eldest or only children are significantly brighter than youngest children; there are marked differences in MCAT scores between those who hope to become nurses and those who hope to become nurses aides, or between those planning to be farmers, engineers, teachers, or physicians; and there are substantial MCAT differences among the various religious groups.

We also tabulated the five principal Protestant religious denominations (Baptist, Episcopal, Lutheran, Methodist, and Presbyterian) against all the other variables, finding highly significant relationships in most instances. For example, only children are nearly twice as likely to be Presbyterian than Baptist in Minnesota, more than half of the Episcopalians “usually like school” but only 45% of Lutherans do, 55% of Presbyterians feel that their grades reflect their abilities as compared to only 47% of Episcopalians, and Episcopalians are more likely to be male whereas Baptists are more likely to be female. Eighty-three percent of Baptist children said that they enjoyed dancing as compared to 68% of Lutheran children. More than twice the proportion of Episcopalians plan to attend an out of state college than is true for Baptists, Lutherans, or Methodists. The proportion of Methodists who plan to become conservationists is nearly twice that for Baptists, whereas the proportion of Baptists who plan to become receptionists is nearly twice that for Episcopalians.

In addition, we tabulated the four principal Lutheran Synods (Missouri, ALC, LCA, and Wisconsin) against the other variables, again finding highly significant relationships in most cases. Thus, 5.9% of Wisconsin Synod children have no siblings as compared to only 3.4% of Missouri Synod children. Fifty-eight percent of ALC Lutherans are involved in playing a musical instrument or singing as compared to 67% of Missouri Synod Lutherans. Eighty percent of Missouri Synod Lutherans belong to school or political clubs as compared to only 71% of LCA Lutherans. Forty-nine percent of ALC Lutherans belong to debate, dramatics, or musical organizations in high school as compared to only 40% of Missouri Synod Lutherans. Thirty-six percent of LCA Lutherans belong to organized non-school youth groups as compared to only 21% of Wisconsin Synod Lutherans. [Preceding text courtesy of D. T. Lykken.]

These relationships are not, I repeat, Type I errors. They are facts about the world, and with N = 57,000 they are pretty stable. Some are theoretically easy to explain, others more difficult, others completely baffling. The “easy” ones have multiple explanations, sometimes competing, usually not. Drawing theories from a pot and associating them whimsically with variable pairs would yield an impressive batch of H_0-refuting “confirmations.”

Another amusing example is the behavior of the items in the 550 items of the MMPI pool with respect to sex. Only 60 items appear on the *Mf* scale, about the same number that were put into the pool with the hope that they would discriminate femininity. It turned out that over half the items in the scale were not put in the pool for that purpose, and of those that were, a bare majority did the job. Scale derivation was based on item analysis of a small group of criterion cases of male homosexual invert syndrome, a significant difference on a rather small *N* of Dr. Starke Hathaway's private patients being then conjoined with the requirement of discriminating between male normals and female normals. When the *N* becomes very large as in the data published by Swenson, Pearson, and Osborne (1973), approximately 25,000 of each sex tested at the Mayo Clinic over a period of years, it turns out that 507 of the 550 items discriminate the sexes. Thus in a heterogeneous item pool we find only 8% of items failing to show a significant difference on the sex dichotomy. The following are sex-discriminators, the male/female differences ranging from a few percentage points to over 30%:[7]

> Sometimes when I am not feeling well I am cross.
> I believe there is a Devil and a Hell in afterlife.
> I think nearly anyone would tell a lie to keep out of trouble.
> Most people make friends because friends are likely to be useful to them.
> I like poetry.
> I like to cook.
> Policemen are usually honest.
> I sometimes tease animals.
> My hands and feet are usually warm enough.
> I think Lincoln was greater than Washington.
> I am certainly lacking in self-confidence.
> Any man who is able and willing to work hard has a good chance of succeeding.

I invite the reader to guess which direction scores "feminine." Given this information, I find some items easy to "explain" by one obvious theory, others have competing plausible explanations, still others are baffling.

Note that we are not dealing here with some source of statistical error (the occurrence of random sampling fluctuations). That source of error is limited by the significance level we choose, just as the probability of Type II error is set by initial choice of the statistical power, based upon a pilot study or other antecedent data concerning an expected average difference. Since in social science everything correlates with everything to some extent, due to complex and obscure causal influences, in considering the crud factor we are

[7] Items reprinted with permision. © University of Minnesota Press.

talking about *real* differences, *real* correlations, *real* trends and patterns for which there is, of course, some true but complicated multivariate causal theory. I am not suggesting that these correlations are fundamentally unexplainable. They would be completely explained if we had the knowledge of Omniscient Jones, which we don't. The point is that we are in the weak situation of corroborating our *particular* substantive theory by showing that *X* and *Y* are "related in a nonchance manner," when our theory is too weak to make a numerical prediction or even (usually) to set up a range of admissible values that would be counted as corroborative.

Some psychologists play down the influence of the ubiquitous crud factor, what David Lykken (1968) calls the "ambient correlational noise" in social science, by saying that we are not in danger of being misled by small differences that show up as significant in gigantic samples. How much that softens the blow of the crud factor's influence depends upon the crud factor's average size in a given research domain, about which neither I nor anybody else has accurate information. *But the notion that the correlation between arbitrarily paired trait variables will be, while not literally zero, of such minuscule size as to be of no importance, is surely wrong*. Everybody knows that there is a set of demographic factors, some understood and others quite mysterious, that correlate quite respectably with a variety of traits. (Socioeconomic status, SES, is the one usually considered, and frequently assumed to be only in the "input" causal role.) The clinical scales of the MMPI were developed by empirical keying against a set of disjunct nosological categories, some of which are phenomenologically and psychodynamically opposite to others. Yet the 45 pairwise correlations of these scales are almost always positive (scale *Ma* provides most of the negatives) and a representative size is in the neighborhood of .35 to .40. The same is true of the scores on the Strong Vocational Interest Blank, where I find an average absolute value correlation close to .40. The malignant influence of so-called "methods covariance" in psychological research that relies upon tasks or tests having certain kinds of behavioral similarities such as questionnaires or ink blots is commonplace and a regular source of concern to clinical and personality psychologists. For further discussion and examples of crud factor size, see Meehl (1990).

In order to further convince the reader that this crud factor problem is nontrivial, let us consider the following hypothetical situation with some plausible numerical values I shall assign. Imagine a huge pot of substantive theories about all sorts of domains in the area of personality. Then imagine a huge pot of variables (test scores, ratings, demographic variables, and so forth) of the kind that soft psychologists have to deal with in nonexperimental work. I remind the reader that I include experimental studies in which these subject-variables, attributes that they bring with them rather than factors we impose by manipulation and randomization, play a critical role in the experimental design as interaction terms. Now suppose we imagine a society of psychologists doing research in this soft area, and each investigator sets his experiments up in a whimsical, irrational manner as follows: First he picks a theory at random out of the theory pot. Then he picks a pair of variables randomly out of the observable variable pot. He then arbitrarily assigns a direction (you understand there is no intrinsic connection of content between the substantive theory and the variables, except once in a while there would be such by coincidence) and says that he is going to test the randomly chosen substantive theory by pretending that it predicts — although in fact it does not, having no intrinsic contentual relation — a positive correlation between randomly chosen observational variables *X* and *Y*.

Now suppose that the crud factor operative in the broad domain were .30, that is, the average correlation between all of the variables pairwise in this domain is .30. This is not sampling error but the true correlation produced by some complex unknown network of genetic and environmental factors. Suppose he divides a normal distribution of subjects at the median and uses all of his cases (which frequently is not what is done, although if properly treated statistically that is not methodologically sinful). Let us take variable *X* as the "input" variable (never mind its causal role). The mean score of the cases in the top half of the distribution will then be at one mean deviation, that is, in standard score terms they will have an average score of .80. Similarly, the subjects in the bottom half of the *X* distribution will have a mean standard score of –.80. So the mean difference in standard score terms between the high and low *X*s, the one "experimental" and the other "control" group, is 1.6. If the regression of output variable *Y* on *X* is approximately linear, this yields an expected difference in standard score terms of .48, so the difference on the arbitrarily defined "output" variable *Y* is in the neighborhood of half a standard deviation.

When the investigator runs a *t* test on these data, what is the probability of achieving a statistically significant result? This depends upon the statistical power function and hence upon the sample size, which varies widely, more in soft psychology because of the nature of the data collection problems than in experimental work. I do not have exact figures, but an informal scanning of several issues of journals in the soft areas of clinical, abnormal, and social gave me a representative value of the number of cases in each of two groups being compared at around $N_1 = N_2 = 37$ (that's a median because of the skewness, sample sizes ranging from a low of 17 in one clinical study to a high of 1,000 in a social survey study). Assuming equal variances, this gives us a standard error of the mean difference of .2357 in sigma-units, so that our *t* is a little over 2.0. The substantive theory in a real life case being almost invariably predictive of a direction (it is hard to know what sort of significance testing we would be doing otherwise), the 5% level of confidence can be legitimately taken as one-tailed and in fact could be criticized if it were not (assuming that the 5% level of confidence is given the usual special magical significance afforded it by social scientists!). The directional 5% level being at 1.65, the expected value of our *t* test in this situation is approximately .35 *t* units from the required significance level. Things being essentially normal for 72 *df*, this gives us a power of detecting a difference of around .64.

However, since in our imagined "experiment" the assignment of direction was random, the probability of detecting a difference in the *predicted direction* (even though in reality this prediction was not mediated by any rational relation of content) is only half of that. Even this conservative power based upon the assumption of a completely random association between the theoretical substance and the pseudopredicted direction should give one pause. We find that the probability of getting a positive result from a theory with no verisimilitude whatsoever, associated in a totally whimsical fashion with a pair of variables picked randomly out of the observational pot, *is one chance in three*! This is quite different from the .05 level that people usually think about. Of course, the reason for this is that the .05 level is based upon strictly holding H_0 if the theory were false. Whereas, because in the social sciences everything is correlated with everything, for epistemic purposes (despite the rigor of the mathematician's tables) the true baseline — if the theory has nothing to do with reality and has only a chance relationship to it (so to speak, "any connection between the theory and the facts is purely coincidental") — is 6 or 7 times as great as the reassuring .05 level upon which the psychologist focuses his mind. If the crud factor in a domain were running around .40, the power function is .86 and the "directional power" for random theory/prediction pairings would be .43.

The division of the statistical power by two on the grounds that the direction of the difference has been whimsically assigned is a distinct over-correction, because the variables found in soft psychology (whether psychometric, demographic, rated, or life history) are by no means as likely to be negatively as positively correlated. The investigator's initial assignment of what might be called the "up" direction, based upon the christening of the factors or observable test scores by the psychological quality thought characteristic of their high end, means that although the theory may have negligible verisimilitude and hence any relationship between it and the facts is coincidental, the investigator's background knowledge, common sense, and intuition — what my friend and former colleague Festinger called the "*bubba factor*" — will cause the predicted direction to be non-random. If we consider, for example, the achievement and ability test area, we have something close to positive manifold, and a knowledgeable investigator might be using a theory that had negligible verisimilitude but nevertheless his choice of *direction* for the correlation between two such tests would almost never be inverse. A similar situation holds for psychopathology, and for many variables in personality measurement that refer to aspects of social competence on the one hand or impairment of interpersonal function (as in mental illness) on the other. Thorndike had a dictum “All good things tend to go together.” I rather imagine that the only domain in which we might expect anything like an evenhanded distribution of positive and negative correlations would be in the measurement of certain political, religious, and social beliefs or sentiments, although even there one has the impression that the labeling of scales has been nonrandom, especially because of the social scientist’s interest in syndromes like authoritarianism and the radical right. There is no point in guesstimating the extent to which the division by two in the text supra is an overcorrection, but it is certainly fair to say that it is excessive when the influence of common sense and the bubba factor is taken into account as a real life departure from the imagined totally random assignment. *The statistical power of significance tests for theories having negligible verisimilitude but receiving a spurious confirmation via the crud factor is underestimated by some unknown but hardly-negligible amount, when we divide the directional power function by two on the random assignment model.*

An epistemological objection to this reasoning holds that it is illegitimate to conceptualize this crazy setup of a pot of theories whose elements are randomly assigned to a pot of variable pairs, since such an hypothetical “population” cannot be precisely defined by the statistician. I cheerfully agree with the premise but not the conclusion. The notion of a random assignment of direction taken as a lower bound to the power, given a certain representative value of the crud factor in a domain, is just as defensible as the way we proceed in computing the probability of poker hands or roulette winnings in games of chance. The point is that *if* we conceive such a class of substantive theories and *if* we assign the theories to the finite but indefinitely large and extendable collection of variables measured in soft psychology (which runs into millions of pairs with existing measures), there is nothing objectionable about employing a mathematical model for generating the probabilities, provided one can make a defensible statement about whether they are lower bounds on the truth, as they are in this case.

If it is objected that the class of experiments, or the class of theories, cannot be enumerated or listed and consequently it is a misuse of the probability calculus to assign numbers to such a vague open-ended class, my answer would be that if that argument is taken strictly as an epistemic point against the application of mathematics to the process of testing theories, it applies with equal force to the application of mathematical methods to the testing of statistical hypotheses, and hence the whole significance testing procedure goes down the drain. When we contemplate the p = .05 level in traditional Fisherian statistics, everyone knows (if he paid attention in his undergraduate statistics class) that this does not refer to the actual physical state of affairs in the event that the drug turns out to be effective or that the rats manifest latent learning, but rather to the expected frequency of finding a difference of a certain size if in reality there is no difference. Assume with the objector that an hypothetical collection of all experiments that all investigators do (on drugs, or rats, or schizophrenics) is not meaningful empirically because of the vagueness of the class and its openness or extensibility. It is *not* defined precisely, the way we can say that the population from which we sampled our twins is all of the twins in the State of Minnesota, or all pupils in the Minneapolis school system who are in school on a given day. It follows that there is no basis for the application of Fisherian statistics in scientific research to begin with. As is well known, were some mischievous statistician or philosopher to say "Well, five times in a hundred you would get this much of a difference even if the rats had learned nothing, so why shouldn't it happen to you?", the only answer is "It could happen to me, but I am not going to assume that I am one of the unlucky scientists in 20 to whom such a thing happens. I am aware, however, that over my research career, if I should perform a thousand significance tests on various kinds of data, and nothing that I researched had any validity to it, I would be able to write around 50 publishable papers based upon that false positive rate."

There are admittedly some deep and fascinating epistemological (and perhaps even ontological?) questions involved here that go beyond the scope of the present paper. All I am concerned to argue is that I will not hear an objection to the whimsical model of random assignment from a psychologist who routinely employs statistical significance tests in an effort to prove substantive theories, since the objection holds equally against that whole procedure. The class of all experiments that will ever be conducted with a choice of a set of variables assigned to a set of theories is no vaguer or more open-ended — and there is certainly nothing self-contradictory about it — than the class of all experiments that people will ever do on the efficacy of penicillin or latent learning in rats, or all of the experiments that all of the scientists in the world will ever conduct, on any subject matter, relying upon statistical significance tests. That fuzzy class is the reference class for the alpha level in Fisherian statistics, so that if the vagueness of such a class of "all experiments using t tests" makes it inadmissible, we have to stop doing significance tests anyway.

7. *Pilot studies*: Low awareness of Cohen's point about inadequate statistical power has been disappointing, but I must now raise a question which Cohen did not discuss (quite properly, as he was not concerned with the

crud factor in his paper): there is a subtle biasing effect in favor of low verisimilitude theories on the part of investigators who take the power function seriously. I have no hard data as to how many psychologists perform pilot studies. A show of hands in professional audiences to whom I have lectured on this subject shows that the practice is well-nigh universal. I do not here refer loosely to a "pilot study" as one mainly oriented to seeing how the apparatus works or whether your subjects are bothered by the instructions or whatever. I mean a study that essentially duplicates the main study, insofar as the variables were manipulated or controlled, the subjects randomized or matched, and the apparatus or instruments employed were those that are going to be subsequently employed. *A true pilot study is, except perhaps for a few minor improvements, a main study in the small*. Such pilot studies are conducted with two aims in mind. First and most often, they attempt to "see whether an appreciable effect seems to exist or not" (the point being that, if in the pilot study one does not detect even a faint trend, let alone a statistically significant difference in the direction the theory predicts, one will not pursue it further). Secondly, for those who take the power function seriously, one attempts to gain a rough notion of the relationship between a mean difference and the approximate variability as the basis for inferring the number of cases that would be necessary, with that difference, to achieve statistical significance at, say, the .05 level. What is the effect of carrying out these two aims via the conducting of pilot studies? A proper pilot study (not merely one to see whether an apparatus works but which, in effect, is a small scale adumbration of the large study that one may subsequently conduct) *is itself a study*. If the investigator drops this line of work as "unpromising" on the grounds that one detects no marginal evidence of an effect, this means that one has conducted a study — perhaps one fairly adequate in design although with small statistical power — and a *real finding* has happened in the world which, under the policy described, never surfaces in the research literature.

Thus, in a given soft area of psychology, or in a subdomain, say, researching a particular variable or instrument, there are hundreds (if you count masters and doctoral dissertation projects, more like thousands) of pilot studies being conducted every year that will never be published, and many of which will not even be written up as an unpublished MA or PhD thesis for the simple reason that they "did not come out right." We do not ordinarily think of this as somehow reprehensible or as showing a bias on anybody's part, and I am not condemning people for this approach, as I have done it myself both in animal and human research. I am merely pointing out that a properly conducted pilot study is itself a research study, and if it "doesn't show an effect" when that effect was allegedly predicted from a certain substantive theory, it is a piece of evidence against the theory. The practice of doing pilot studies, not to publish them but to decide whether to pursue a certain line, means that there is a massive collection of data, some unknown fraction of which comes out adverse to theories, that never sees the light of day. *A fact that occurs in the laboratory or exists in the clinic files is just as much a fact whether Jones elects to write it up for a refereed journal or to forget it*. A class of potential falsifiers of substantive theories is subject to systematic suppression in perfectly good scientific conscience. We do not know how big it is, but nobody knowledgeable about academia could conceivably say that it was insignificant in size.

As to the second function of pilot studies, if, as above, we conceptualize the subset of substantive theories that have negligible verisimilitude linked randomly to observed variable pairs so that the "relationship" between the theory and the facts is coincidental, such investigations are dependent upon the crud factor for obtaining positive results. Now what will be the effect of the almost universal practice of doing pilot studies in such a situation? If the crud factor in a subdomain is large, investigators doing pilot studies on it will get positive results and proceed to do the main study and get positive results there as well, although they will not have to expend as much energy collecting as many cases as they would if the crud factor in the subdomain were small. But given that the crud factor, even if small, is not zero in any domain, investigators operating in an area will discover (on the average) that an effect exists, the crud factor being ubiquitous. Depending upon their motivation and the kind of data involved (e.g., questionnaires are easier to collect than tachistoscope runs), they will see to it that the main investigation uses a large enough sample that statistical significance will be achieved by differences roughly of the size that they found in the pilot study. Hence, investigators eager to research a certain theory in a loosely specified fact domain *are not wholly at the mercy of the crud factor size*. They are not stuck willy nilly with a fixed power on its average value in the domain, since the second (and legitimate!) use of pilot studies is precisely to see to it that a trend of roughly so and so size, if found in the pilot study, will be able to squeak through with statistical significance in the main study subsequently conducted. The limiting case of this, which fortunately does not exist because money and time are finite, would be that of investigators doggedly pursuing the testing of a particular theory by regularly doing pilot studies that give them a fairly accurate estimate of the size of the trend and if a trend seems to exist, invariably conducting the large scale experiment with a sample size having nearly perfect statistical power. This nightmare case is therapeutic for psychologists to contemplate, if as I believe, *it is the present real situation writ large and horrible*. It means that in the extreme instance of a pot of theories none of which has any verisimilitude, they would all come out as well corroborated if the investigators consistently

(a) did not pursue those lines where the pilot studies suggest a trend in the opposite direction and hence "not a profitable line to follow up" and, on the other hand, for those that are in the theoretically expected direction, (b) invariably did the main experiment using huge *N*s generating super power. I do not know, and neither does anybody else, how much different the present situation in the social sciences is from that nightmare situation, especially given the terrible publish or perish pressure upon young academics.

8. *Selective bias in submitting reports*: Some years ago, when the Minnesota Psychology Department moved to a new building, I went through my old files of research studies and discarded those that it was obvious I was for one reason or another never going to submit for publication or pursue further. Due to my Minnesota training by Hathaway, Paterson, and Skinner, all of whom (for quite different reasons) disparaged the publication of piddling average results with large overlap merely because they achieved statistical significance, I found a number of studies that were statistically significant that I had not submitted. This was mainly in the clinical area, and what it meant was that a particular MMPI scale or a currently popular test for detecting minimal brain damage was only "statistically significant" but not of practical value because of large overlap. But, I did find quite a few studies of adequate sample size so that one could trust the results from the power function standpoint that were not submitted simply because they did not show a trend. For technological purposes this may be legitimate, although I am inclined to think not. For purposes of testing substantive theories, it is an example of bias. Several rat latent learning studies that I had done with MacCorquodale were not submitted because they were in the grey region, showing probabilities hovering around .10, so that we were hardly in a position to say much as to whether the rats had learned anything or not.

I began inquiring among colleagues and students as to whether they thought they were completely even-handed in submitting to journals results that achieved statistical significance and those that did not, and I haven't found a single person who claims to have been. Some people say rather shamefacedly that they are not even-handed in the matter, and others point out — sometimes quoting R. A. Fisher on the point — that a statistically significant finding "proves something" whereas a failure to refute H_0 does not prove anything. This of course involves a serious disarticulation between Fisherian statistics and Popperian ideas of theory testing, since if a null result does not disprove anything it is at least arguable that one ought not to have done the experiment in the first place, because he is only going to count it if it comes out "positive," a major Popperian sin! I do not want to go into the very difficult and technical philosophical issues of that problem, but the point is that everyone acknowledges, some more freely than others and some by offering justificatory arguments, that they are considerably more likely to submit an article to a journal if they got significant results than if results failed to reach significance. This stems partly from thinking that a null result “doesn’t prove much,” partly from a recognition of the role of inadequate power, and partly because of the next factor to be discussed, to wit, that editors, also recognizing an epistemic asymmetry here, are more likely to reject a paper that fails to reach statistical significance, especially if it is plausible to attribute it to inadequate power.

9. *Selective editorial bias*: I find no hard data in the literature on the practice of editors and referees, but people who have served in either of these capacities report the same thing as investigators do: they are somewhat more favorably disposed to a clear finding of refuted H_0 than one that simply fails to show a trend. And here, as in the case of investigators’ bias, this behavior can be defended on the ground of what R. A. Fisher said about asymmetry.

10. *Detached validation claim for psychometric instruments*: It is typical of research articles in soft psychology that a certain instrument, say, the MMPI social introversion scale, is going to be employed to test a substantive theory in which a concept claimed to be measured by that instrument is one of the embedded constructs. The investigator accepts an obligation to persuade the reader that the scale is valid for the trait (attribute, construct) that it names. Whether one can simultaneously validate a psychometric instrument and corroborate a substantive theory in which the construct labeled by the instrument’s title finds a postulated place is a deep matter which was discussed over 30 years ago by Cronbach and myself (Cronbach & Meehl, 1955), and I shall say no more about it here. But it would seem that, if I am going to confirm or refute a theory about the relation of social introversion to anxiety-based affiliative drives, I ought to have grounds for thinking that the test is sufficiently valid for use in the way I am using it, since the internal network of most experiments is not sufficiently rich to make a strong argument of the kind that Cronbach and I offered in 1955 about simultaneous testing.

How do investigators typically go about buttressing this initial validation claim for an instrument so that they can get on with doing the main study? We all know how it looks in the journals. What the writer does is to list a series of authors who have either “validated” or “failed to validate” the introversion test and perhaps summarizes by counting noses as to how many found it had validity and how many did not and then, if we are lucky, gives us a representative validity coefficient. What happens next? The author writes some such sentence as, “Since the majority of studies showed respectable validity for Fisbee’s Test, and in the more favorable studies the validity coefficients were in the range .40 to .50, it was felt that Fisbee’s Test was perhaps the best available test of introversion and it was therefore used in the present investigation." Following this summary remark, a *qualitative* claim of "respectable validity" and a *quantitative* claim based upon a representative value, the rest of the article typically presupposes that X is valid for introversion without any further reference to the actual numbers. Typically, negligible attention is given the unreliable component as pushing correlations down and the reliable but invalid component of the test score as pushing them sometimes up but possibly down (since we cannot be completely confident about the *ceteris paribus* clause). Except for persons strongly interested in methodology and concerned with the problem presented in Campbell and Fiske's classic paper (1959) and its relationship to factor analysis, I find few psychologists who are sensitive to this obfuscating factor in the qualitative sense, let alone appreciative of its likely quantitative influence. This blindness comes, I think, from the yes-or-no use of the word "valid" (or "validated"), which is a bad semantic habit acquired in beginning psychology courses due to the somewhat crude and inaccurate way the concepts of reliability and validity are typically presented. Every sophisticated psychologist, and certainly anybody concerned with psychometric theory, knows that tests have multiple validities or, putting it another way, they have a validity and multiple invalidities, depending upon which component you have your eye. But the verbal habit of saying that a test "has been validated," and hence, for purposes of the current research we are engaged in, "can be taken as substantially valid," prevents an adequate appreciation of the danger of a detached validity claim.

Having shown by a survey of studies that validity coefficients are in some range we agree to consider respectable (and I cannot resist pointing out that validities of .40, accounting for one-sixth of the variance, would hardly be considered "respectable" by a chemist or geneticist), the nonquantitative blanket category word "valid" is now picked up and employed in the subsequent discussion in the research paper. Obviously this can present a very misleading picture if the reader does not keep harking back to that distribution of validity coefficients which are *not* subsequently mentioned.

In deductive logic one speaks of the Rule of Detachment, which says that if we have written $p \rightarrow q$ and we have also written p, then we are entitled in the rest of our discourse to assert q without having continually to repeat the syllogism or allude to the process of deductive inference. In that context all is well, and I have spoken of 'detached validation claim' to highlight the point that *in the inductive logic of a pattern of correlations of various sizes, nothing comparable to a rule of detachment can properly operate with respect to the qualitative designation "valid."* This is so obvious a point that its statement would seem to suffice, and I do not expect any reader to disagree with me about it. But, is it a point of any quantitative impact on interpretation? It most certainly is, as is shown by the following numerical example employing values not the least outlandish as data go in soft psychology.

Suppose I am going to investigate some theory about introversion and intend to rely on the *Si* scale of the MMPI for measurement of an individual differences component that's going to interact with some experimental factor in my design. I offer a representative validity coefficient of say .40 (not unlike what we find, and one can find examples in the literature of soft psychology where people invoke validities less than .40 as "at least substantial validity for trait X" in an instrument they wish to use in their experiment). Suppose I properly report that I have a reliability coefficient of say .80 for my introversion test. It is interesting how happy psychologists are with high reliabilities and low validities, since it should occur to one that this combination might have unfortunate consequences due to the possible falsity of the *ceteris paribus* clause. Sixty-four percent of the variance of the observed test scores is reliable variance and 16% is valid variance, i.e., valid for my purposes. Hence about three-fourths of the reliable variance is invalid variance, that is, it is measuring something non-chance but not what we have our eye on, not what the scale is named for. Absent further data which may or may not be known and which in any case is likely not to be reported by the investigator and hence not available for the reviewer of research in a domain, one does not know whether that three-fourths of invalid reliable variance is a collection of other smaller factors or possibly even one or more factors that may be larger than the reliable component named by the test! Now when I find out that such a test correlates significantly with something else in my design, with what confidence am I entitled to attribute that to introversion, when three-fourths of the reliable variance of the test is something other than introversion as I have conceptualized it? Or putting it the other way around, what about the possibility that the three-fourths of reliable but invalid variance counteracts the influence of the validly-measured introversion component in my particular design; consequently the falsity of the *ceteris paribus* clause with respect to components of the invalid variance prevents me from achieving a significant result even if my theory about introversion and its interaction with a manipulated variable has verisimilitude? It may be objected that it would be too onerous to require that investigators plug in a whole bunch of things that they ought to be worried about with the Campbell-Fiske discriminant validation in mind. All I can say to that is that, absent a tradition of so doing, I do not know how much confidence to have in detached validity claims for testing substantive theories.

This concludes my list of obfuscating factors. I hope I have convinced the reader that they are almost always if not invariably present in a soft psychology study, and that their quantitative impact, while varying from one project to another and on the average from one domain to another, will rarely be of negligible size. The worst part of it is that obfuscating factors (1) to (5) will tend to make good theories look bad; obfuscators (6) to (9) will tend to make poor theories look good; and I suppose obfuscator (10) can work either way and we frequently would not know in a given research design which way it might be working. As I said in my introductory thesis, we have ten factors, sizeable, variable, and typically unassessed in a given setting, working in opposite directions to produce a net "box score" of successful or unsuccessful attempts to refute the null hypothesis.

It might be objected that, in assessing the state of the empirical literature on a theory, we take it for granted that "all theories are lies," that it is not a question of a theory being absolutely true or totally false but that some theories are in better shape than others, and what the reviewer does is evaluate the evidence in some over-all sense. It is true that reviewers sometimes focus on particular assumptions of a research study or of a subset of studies all employing the same basic design or instrument, but this focusing is not very helpful in adjudicating the merits of the theory at the end of the review article unless it leads to some sort of *refined* "box score" from which highly doubtful studies, excessively problematic because of their auxiliary assumptions or because of low statistical power, are excluded. Ideally, this would be what happens, but since the ten obfuscators vary in size and in most cases we do not have a rational basis for assigning a numerical value to an obfuscator's influence one way or the other (e.g., can the reviewer assign a numerical value to the probability of a specific auxiliary?), the box score is going to be slanted one way or the other very much depending upon the reviewer's crude sifting on the basis of commonsensical or theoretically-based assignment of these unknown numerical values.

I do not believe psychologists in the soft areas can take much consolation from what is admittedly a correct statement in rough qualitative form, "We don't allow a single experiment however replicable to kill a theory that is otherwise good and we aren't impressed with a collection of truly feeble statistical significances, rather we look at the over-all shape of the factual terrain and make a reasoned judgment." I do not suggest that there ought to be an inductive algorithm which avoids the necessity for making reasoned judgments, although because of some of my writings on clinical decision making people have attributed that idea to me. My point is rather that I don't see how cogently "reasoned" a so called "reasoned judgment" can be when faced with ten obfuscators, none of which is likely to be of negligible influence in a given research domain, and some of which may be of very strong influence in one subdomain or another, varying a lot from one subdomain to the other, most of them not accurately assessable even in a subdomain let alone in a particular research study reported on, and operating in opposite directions.

Consider a domain in which only two auxiliaries are needed to make the derivation to the predicted empirical result and both are stated explicitly, say, an "input" auxiliary which postulates that a certain experimental manipulation will induce such and such an inferred psychic state in subjects, as in our introversion example above, and only a single "output" auxiliary, such as a piece of psychometric theory about the Rorschach or MMPI. Surely this is a conservative case since, if spelled out in the way we don't because of obfuscator (1), there are likely to be several input and output auxiliaries in most empirical tests of a theory that possesses sufficient conceptual richness to be causally interesting.

Hull and Co.'s famous *Mathematico-deductive theory of rote learning* (1940) is seldom read today, even for historical interest, because the theory, a kind of tour de force, is pretty well dead. Some poked fun at Hull and his collaborators for going through all of that symbolic logic (one of the collaborators was a mathematical logician brought into the crew specifically for that purpose) as not really necessary. Some critics even called the use of logic "mere window dressing" because Hull was infatuated with philosophy of science. But one thing that book made very clear, which was *not* clear to everybody before, was the fact that when one really requires one's derivation chain to be deductive and rigorous, it turns out one has to put in an awful lot of statements which a nonlogician psychologist would either take for granted or would not even be aware were required to make the derivation. And if such a thing is true in the case of a theory about memorizing nonsense syllables, *a fortiori* it is true in theories about emotion, motivation, social perception, achievement, and the like.

Consider a domain in which the true theory involves a quantitatively moderate effect that would amount to a Pearson correlation of .50 between the observables. Suppose both the input and output auxiliaries have a probability of .85, the *ceteris paribus* clause is moderately dangerous, say its probability is .80, and the experimenter's faithful fulfillment of the conditions as described is .90. The expected value of the mean difference on the output variable cutting the input variable at its median as described above is .80 standard score units, which is 1.74 *t* units from the value required for a 5% directional significance test, so we have a power of .96. Multiplying the power by the product of the two auxiliaries and of the *ceteris paribus* and experimental conditions gives us a net probability of a "successful" outcome of .59. So of 100 studies conducted in such a domain, we might expect around 59 to come out "positive," meaning a statistically significant result in the predicted direction, and the other 41 to come out negative, that is, adverse to the theory. Suppose that there is a strong bias in submission and editorial acceptance in that all positive studies are submitted and accepted (absent a gross defect in design or analysis which I will for the moment assume is not happening), but only a bare majority of negative studies are submitted. Of those submitted, a bare majority are accepted by the editor. So all 59 positive studies are submitted, and all those submitted are accepted, whereas of the 41 negative studies 11 are submitted and accepted. Hence the box score for true theories is 59/(59 + 11) or 84% "successes" *in the literature*.

Compare that with the case of a wholly false theory, that is, one having zero verisimilitude. While the crud factor in this domain is only .30 or even .25, the use of pilot studies to decide whether to pursue it further results in an effective crud factor in the domain somewhat higher, say, of .40 in those variable pairs that are further pursued by investigators who do not report the pilot studies on the pairs they drop. Then the expected value of the mean standard score difference on the output variable will be (.40)(1.60) = .64 which is 1.07 *t* units from the critical *t*, yielding a power at the 5% level of .86. If now we assume (neglecting bubba factor, common sense, and the tendency for correlations in the randomly chosen variable pot to be positive), a "pure chance" situation as to directionality, we must divide that power by two, getting a power of .43 on a wholly random pairing. Since this theory has no verisimilitude, the auxiliaries and derivation chain are irrelevant. What we are dealing with is a crud factor probability of a positive result. So of 100 studies conducted 43 come out positive and 57 come out negative. Again, all 43 positive ones are submitted and if their design is otherwise adequate are accepted, whereas of the 57 negative studies (.51)(.51)(57) = 15 are accepted. So the box score for this totally false theory is 43/(43 + 15) or 74% "successful" outcomes. On these perhaps somewhat pessimistic but not farfetched assumptions about the domain, *true theories and false theories show box scores in the literature that are only about 10% different from each other*. Surely we cannot suppose that a sympathetic but skeptical reader can interpret *Bulletin* articles meaningfully, realizing that such a domain situation is possible and not wildly improbable? If the reader will plug in some other values he will, I think, be impressed with how wildly the box score percentages can bounce around as a function of trustworthiness of auxiliaries and the extent to which the use of pilot studies has led to an exclusion of those variables whose crud factor is low. Without making outlandish assumptions, one can show that in one domain the box score for theories with zero verisimilitude could run higher than for a perfectly true theory in some other domain.

One of the biggest contributions to this frightening possibility is in the economics and sociology of science. Differences in availability of money for currently popular fads being studied by highly visible psychologists, and the pronounced differences among theories and domains with respect to the ease of increasing statistical power by boosting *N*, will mean that a prestigious investigator, who has an easy time getting a grant and whose method of study is questionnaires, is going to get a lot of mileage out of the crud factor compared with a graduate student, little known investigator, or someone working in a domain not currently popular and whose data are of an experimental nature or involve extensive testing of individuals, so that the purely logistical and temporal difficulties of accumulating a large *df* mean that the researcher will have a lot more trouble eking out statistical significance on the basis of the crud factor. The investigator may have so much trouble reaching statistical significance even with a high verisimilitude theory that the expected value of the box score is actually lower than for a false theory in a domain of the other sort.

These are not far-out, nightmare, implausible occurrences. I am not relying on the fact that the statistician tells us in advance that once in a while we will be committing a Type I error, which is not the point at all. Type I errors in the mathematical sense have not been adduced at any point in this paper and will not be. We could add insult to injury by including considerations about the prior probability of substantive theories in soft psychology which — if one goes by the track record of history — must be considerably less than one half and, as I read the record, would be running down around 10%, if that high. For example, in my youth there were a half dozen major theories of animal learning (Hull, Guthrie, Tolman, Skinner, etc.) and a few minor ones, all of which I think it fair to say have been refuted, although some are capable of covering more of the fact domain than others. If the track record for theories of rat learning suggests a prior probability of truth (or of verisimilitude high enough to remain in the running after a generation of research) as low as .16, I cannot get myself to believe that the corresponding prior probability for theories in personology, psychodynamics, or social psychology is higher than that. If the likelihood ratio of a theory on its evidence at a given time were as high as 2 to 1 (based on the conditional probabilities given my obfuscating factors), but we take into account that the prior on any given theory for either truth or very high verisimilitude is, say, only one in 10, then the Bayes Formula posterior probability on the theory having high verisimilitude is still only .20, so that the odds are still running 4 to 1 against! But since many psychologists seem to think that the prior on theories in soft psychology is pretty good — I cannot for the life of me understand why they think this, either from armchair grounds or from the track record of our field — I will forego further discussion of that aspect of the problem. One does not have to be a Bayesian in one's view of statistical inference to accept the statistical reasoning on which this paper relies.

Students and colleagues sometimes respond to these pessimistic notions by saying, in effect, "Well, I don't know *exactly* what's wrong with the reasoning of Bakan, Lykken, Meehl, Rozeboom, and Co. (Bakan, 1966; Lykken, 1968;

Meehl, 1967, 1978; Rozeboom, 1960) but it's obvious that there must be something wrong with it, because significance testing has worked fine in agronomy, which is where R. A. Fisher developed most of it." I do not know whether Sir Ronald was impressed by the progress of theories in psychology and sociology. I have been told that he looked upon them with considerable disdain, but I am unaware of any published statements to this effect. The reader should not reassure himself by *ad verecundiam* in the name of the great R. A. Fisher, with whom I am not in any kind of technical mathematical combat (a combat I would be certain to lose). Assuming it true that significance testing enabled great strides to be made in agronomy (I am acquainted with a biometrician who has doubts on that score which I am not competent to assess), this cannot provide reassurance with regard to my ten obfuscators in testing theories in soft psychology because there are several differences between the two domains. These differences are intimately connected, but they do represent different ways of looking at the problem so I will distinguish them without pressing the possibility that they can be reduced to one core difference. That might not be persuasive to some readers, and there is no harm in separating them even if they do have a deep common root.

The first difference is that investigating whether manure is better than potash for fertilizing corn is essentially a technological question rather than the testing of a substantive theory, unless the term 'theory' is used in a broader sense than that which this paper is about. The efficacy of a fertilizer on plant growth is a question similar to a comparison of two sulfonamides in the treatment of strep throat, or the question asked a quality control statistician when he is requested to determine at a fixed confidence level whether more than 2% of the cartridges manufactured by an ammunition factory are defective.

Second, experimenters in agronomy develop a somewhat implicit lore about the subject matter including a rough range of economically and logistically feasible values of the manipulated variable, as well as plausible empirical bounds on the output increments. Thus, no one proposes to apply potassium nitrate in a density of a pound per square yard, and no one expects to quintuple the yield of wheat from any economically feasible amount of fertilizer. Even a statistician who, due to strong Fisherian identifications, has a distaste for the decision theoretical term 'power,' has a pretty good idea of the number of plots in a design of a certain logical complexity that is likely to be needed to detect a difference of the size that the agronomist cares about as worthwhile. If we get a 2% increase in wheat yield using fertilizer F_1 over fertilizer F_2 when the more effective one is 20% more expensive to the farmer, we are not going to fool around with such a thing. A comparable "reasonable range" of either input or output usually does not exist, or at least is not as narrowly demarcated, in the case of testing theories in soft psychology. This problem of the selection of appropriate levels of experimental factors remains an unsolved problem of psychological methodology despite the important methodological contribution of Brunswik (1947) concerned with representative design. The great majority of investigators in theoretical psychology pay very little attention to Brunswik's powerful arguments, so that after all these years most investigators will focus all of their planning concerning representativeness on that of the sample of organisms, negligible attention being paid to representativeness or stratification of the experimental factors, whether manipulated or differential.

Third, my thesis concerns nonmanipulated factors either as main influences or as potentiators of a manipulated factor in an interaction effect. Agronomy deals with *experiments*, not correlational studies of purely cross-sectional data. Even the corresponding variable, the "individual differences" variable, which belongs one might say to the micro-regions of soil or to the grains of wheat seed, is in agronomy not quite like psychology because wheat strains can and will be chosen for *appropriate economic inference* after the experiment is done, which need not involve any problem of representativeness of design.

Fourth, and this is the most interesting methodologically as I have pointed out elsewhere (Meehl, 1978), *there is a negligible difference between the substantive theory of interest and the counter null hypothesis in agronomy, whereas in theoretical soft psychology they are distinctly different and frequently separated by what one could call a large "logical distance*." If I am testing Festinger's theory of dissonance or Meehl's theory of schizoidia or Freud's theory of dreams by a correlational study in soft psychology, the propositions of the substantive theory, even taken jointly with their implications, are not the logical equivalent of the statistical hypothesis of a directional difference which I attempt to prove by refuting a directional null hypothesis. The unfortunate conflation of these two things in statistics courses, in which the word 'hypothesis' is used throughout as if one did not have to worry about this critical distinction, leads the psychologist who does not reflect upon the epistemology of the situation to think of them as nearly the same, although very few would maintain that error on reflection. The psychologist often does something he has been taught not to do in the statistics course, namely, he thinks of the "opposite" or "alternative" to the null hypothesis as somehow constituting the hypothesis he is testing, and as a result he is tempted to think (despite the undergraduate statistics class warnings) that if the *t* test, *F* test, χ^2 or whatever has a probability only .05 of arising on the null hypothesis, then it's "sort of true" that he can be 95% confident that the alternative — which he then translates as the directional difference — is true. Then, because he does not distinguish theory and hypothesis clearly, it seems (vaguely) as if he can be 95% confident that his substantive theory is true. Nobody who got an "A" in a statistics course is likely to make the first of these mistakes, although one frequently runs across persons who can be seduced into saying something close to that on a PhD oral. But even if he avoids making the first of those mistakes, or tries to legitimize it on some Bayesian ground of which Fisher would not approve, he still may attach the confidence level to the substantive theory. The point is that one does not have to make an explicit mistake in undergraduate statistical formulation to make a more subtle mistake of thinking (roughly and inexplicitly) that somehow if the probability of getting what we got is very small if there were no difference, then we can be quite confident that there is a difference, and then we equate the existence of a difference with the theory that suggested the difference to us. Consequently, without exactly taking the complement to the significance level as our theory-confidence, we nevertheless think it must be "quite large," as long as the significance level we have achieved is "quite small."

I am convinced that both among students and faculty this inexplicit, surreptitious carry-over of a confidence, of a strength of belief in the substantive theory because it is vaguely associated in one's mind with the statistical hypothesis that is considered the alternative to the directional H_0, is quite common. I suggest this is only partly because of the fact that statistics books and lecturers in elementary statistics use the word 'hypothesis' in a somewhat indiscriminate way, not highlighting the difference between a substantive (causal, structural, or compositional) theory and a statistical hypothesis about numerical values of observables. It is also because only a minority of social scientists ever take a course in either philosophy of science or freshman logic, so they don't get exposed to the logician's business about inductive inference being an invalid syllogistic figure. As everyone learns in beginning logic, while *modus ponens* and *modus tollens* are valid syllogistic figures, what used to be called the ordinary "confirmation" hypothetical syllogism $p \rightarrow q$, q, $\therefore$ p is, alas, deductively invalid and at first blush appears to be the form of inference in empirical science. As one of the logic texts I studied as an undergraduate neatly put it, "Elementary logic books are divided into two parts. In the first part, on deductive inference, the formal fallacies are explained; in the second half, on inductive inference, they are committed." To go into the current state of confirmation theory is beyond the scope of this paper and my competence. However, when social scientists are not sufficiently alerted to the elementary logic that the inductive inference is a formally invalid figure, they sometimes talk as if there were some kind of solid gold *proof* possible using an inductive inference, of a kind which philosophers agree cannot exist. That this is not an imaginary danger is shown by the frequency with which criticisms of research studies that conclude for a certain theory contain sentences like “But merely because *X* correlates with *Y* does not prove that …” which, given the nature of inductive logic, is a trivial remark unless expanded in a form that explains why such and such an observational result does not tend strongly to *confirm* or *corroborate* a substantive theory, which is presumably what the writer wanted to say and would have said had he been more sophisticated in the logician’s terminology.

But, it may be objected, we were supposed to be explaining why theory-testing by null hypothesis refutation in soft psychology may be a rather weak and misleading strategy despite the success of that approach in agronomy. Don’t these troubles about formal logic and the inherent fallibility of all inductive inference apply equally strongly there? The rebuttal to this objection takes us to the heart of my doctrine in this paper. While all ten of the obfuscators play important roles in causing trouble for the investigator of a psychological theory, and while the fact that they are so numerous, variable, and countervailing makes the task of unscrambling well-nigh hopeless in some domains, this point about the logical distance between statistical hypothesis and substantive theory, *when combined with the crud factor*, introduces a difference between correlational theory testing in soft psychology and experimental manipulation in agronomy that amounts to a difference of kind and not of degree. If I don’t manage to convince readers of anything else, I will have succeeded in large part if I convince them of this radical qualitative difference. It is precisely the logical distance between the statistical hypothesis and the substantive theory, when combined with the ubiquity of nonzero correlations, that makes current strategy radically defective and probably not improvable, even if the other obfuscators could be eliminated or greatly reduced in their size and influence.

When one has distinguished clearly a substantive theory from a statistical hypothesis (*which in agronomy is an “hypothesis” subject to problematic induction only because of sampling error — not* because the subject matter is about hypothetical constructs or unobserved events in the past, as in psychology) both in his concepts and his semantic habits, one sees the following point immediately: Suppose our null hypothesis in agronomy (or medical testing, or quality control, or any of those minimally theoretical, mainly technological domains to which statistics is applied) is that “potash makes no difference to wheat yield,” or “tetracycline makes no difference to strep throat,” or “there are not more than l/10th of 1% defective cartridges in this batch.” We neglect the possibility that potash or tetracycline has an adverse effect. (If preferred, reformulate the null hypothesis as a directional null hypothesis to the effect that “potash either has no effect on the growth of corn or affects it adversely.”) Despite the vagueness of the directional null, including not merely H_0: d = 0 but everything on the wrong side of it (which led Fisher not to like this form), it is a matter of logic, independent of one's statistical orientation or the power function or anything else, that the directional null and its counter null exhaust the possibilities. If the directional null "potash has no effect or an adverse effect" is false, it follows as the night the day that the counter null "potash has a positive effect" must be true.

But, you may say, I have surreptitiously shifted from a statistical to a causal statement. Yes, so I have, because it is so easy *and harmless* in this instance. Between the statement "plots of corn fertilized with potash differ from plots of corn not fertilized with potash" stated in purely statistical terms without reference to causation, and the substantive "theory" of interest, that "putting potash on plots of corn seeds increases the yield of corn," is not a difference that anybody but a philosopher cares about. It's not a difference that makes a difference. Even a philosopher, if he is a philosopher of science talking about methodology, would allow himself to move freely back and forth between the statistical counter null and the causal substantive "theory" that potash helps one grow more corn. *Except in a seminar on Hume, nobody bothers to distinguish between the counter null hypothesis and the causal conjecture in agronomy*. Not a farmer, a professor of agricultural economics, the sales director of a fertilizer manufacturing company, or a politician in India cares one whit about the fine-line distinction between "fertilized plots have a bigger yield" and "fertilizer produces a bigger yield." The nature of the problem and our general background knowledge guarantee that there will be no difference between these two that's worth talking about unless you were discussing Hume and the metaphysics of causality in a philosophy seminar. For example, nobody in his right mind thinks that harvesting corn in the late summer exerts backward causality upon what we did in the spring about fertilizing plots, let alone that the process was based upon a table of random numbers! Plants take the substance they use in growing from the air and soil; for heaven's sake, where else would they get it from? Long before modern biochemistry, every farmer, going back to thousands of years B.C., became unavoidably aware of the fact that some soil was "better soil" than others for growing purposes. Whether we start with the background knowledge of horticulturists, botanists, and biochemists, or the background knowledge of my sainted grandmother who never finished the third grade, we *know* that plants get their nutriment from the soil. Today we also know scientifically about the fixation of nitrogen, etc. If we put chemical compounds or animal products that contain nitrate radicals that can go into solution into the ground, it does not take a PhD in physical chemistry to figure out that this might be a plausible way for plants to grow better.

But I don't want to engage in overkill. The simple and obvious point is that there is no appreciable difference between the semantic content of the counter null hypothesis (proven with the same confidence with which we have refuted the directional null) and the "substantive theory" that there is a causal connection between fertilizer and yield. If a certain fertilizer has no effect on a certain type of plant, or if two different fertilizers have an equal effect, the null hypothesis will be literally true, and we will correctly fail to refute it except for Type I errors in 5% of the cases if that is our alpha level. Neither of these two things obtains in soft psychology. There is a vast difference, involving numerous intervening steps and auxiliary assumptions, between "Meehl's theory of schizotaxia is substantially correct" and "Many schizophrenics show a ± dysdiadochokinesia." Secondly, the improbability of statistically refuting H_0 set at some high significance level is equal to that significance level when H_0 is literally true, but the improbability of successfully refuting it at that same level is much different in a domain where everything is correlated and the crud factor is not of negligible size.

Another way of looking at the problem is in terms of existing competitor theories, some formulated, some easily formulable with a little imagination. For most statistical findings in soft psychology studies, I daresay a group of faculty or graduate students could come up with a dozen plausible alternatives to the theory of interest if allowed a morning's conversation over coffee and Danish, whereas in the agronomy case there are no such plausible alternatives. If somebody in agronomy were to say, "Well, since you have refuted the null hypothesis at such a small alpha level, and several other people have replicated your result, I grant that fertilized plots in England, India, and Iowa yield more corn. But that doesn't conclusively *prove* (= *demonstrate*, deductively) that the fertilizer had anything to do with it," the obvious reply would be an incredulous "Oh, strictly speaking we haven't a deduction, but what in the devil else would you have in mind?", and to this counter question no sane option would be forthcoming in the agronomy case. If that same counter question were put in discussing Meehl's theory of schizotaxia, Festinger's theory of cognitive dissonance, Freud's theory of dreams, Schachter's theory of affiliation, or other theories in soft psychology, it would not be difficult for the questioner to come up with alternatives. Even if one did not have enough imagination or smarts to come up with plausible looking alternatives, he could always say simply, "Well, there are always alternative explanations of anything complicated, we take that for granted in science and in philosophy, do we not?" Meehl, Festinger, Schachter or whoever would have to say "yes" to that, whether or not the questioner was motivated and ingenious about inventing specific competing theories.

Despite the current technical problems in confirmation theory among philosophers of science, there is nothing obscure or recondite about the point I am making here. It is not commonly seen because of the way null hypothesis testing is taught in statistics courses, but it is not difficult to see. If I refute a directional null hypothesis in agronomy or in a biochemical medical treatment, I thereby prove (in a strong although not strictly deductive sense of that term) the counter null; and the counter null is essentially equivalent to the substantive theory of interest, namely, that fertilizer makes a difference to corn, or tetracycline to strep throats. If you have “almost conclusively proved” the one, you have “almost conclusively proved” the other. But a complex substantive theory involving hypothetical psychological entities, states, and processes, conjectured residues of past learnings in the life history, latent contents underlying dreams or parapraxes, “factors” influencing the correlations of psychometric instruments — here it is not strong “proof” of anything to refute either the point or the directional H_0, because of the crud factor. So that whatever theory we happen to be talking about, we know that the correlations will *not* be zero and that we will *show* them not to be zero, given sufficient statistical power. Hence, (nearly) definitive falsification of the directional null hypothesis, while it (nearly) conclusively proves the directional counter null (taken literally, a trivial result given the crud factor, except for the directionality), does *not* thereby prove with high confidence the truth of the substantive theory. *The substantive theory has a host of alternatives*, some of which are interesting theoretically, some of which are not, and *most of which nobody has thought of* but could in a morning’s free-wheeling speculation. That is simply not the case in agronomy, or the testing of the therapeutic efficacy of a drug, or sampling from a batch of rifle cartridges.

I believe that the foregoing line of argument, although it may be subject to some degree of quantitative correction here and there, is unanswerable. I have been teaching it to classes of doctoral candidates for 20 years, aided and abetted by a couple of Bayesian statisticians who come in as guest lecturers, and I have not heard a strong objection, reply, or “effective softening of the blow” yet. Nor have I heard such from colleagues with whom I have conversed or corresponded. I am inclined to think that if 300 doctoral candidates at a first-rate psychology department, not to mention several PhD candidates in statistics, psychometrics, and philosophy of science who have taken the course, and perhaps two dozen eminent psychologists who have been exposed to these ideas in similar form, have not come up with an answer to this line of thought, then if not substantially correct it must contain a mistake of great depth and subtlety. In what follows I shall therefore allow myself the assumption that, pending better instruction and until further notice, I am correct in viewing these ten obfuscators as strong, variable, countervailing, and from case to case not accurately estimated, supporting my thesis that: *Null hypothesis testing of correlational predictions from weak substantive theories in soft psychology is subject to the influence of ten obfuscating* *factors whose effects are usually (1) sizeable, (2) opposed, (3) variable, and (4) unknown. The net epistemic effect of these ten obfuscating influences is that the usual research literature review is well-nigh uninterpretable.*

I do not subscribe to the pollyanna doctrine that one should not engage in "purely destructive criticism" if he doesn't have anything to offer instead. There is such a thing as killing a theory even though one is not prepared to advocate another one, although admittedly the ideal Popperian case is two theories in competition which are sufficiently strong that the corroboration of one theory by a risky point prediction involves observing a numerical value that slays the other theory *modus tollens*. I am prepared to argue that a tremendous amount of taxpayer money goes down the drain in research that pseudotests theories in soft psychology and that it would be a material social advance as well as a reduction in what Lakatos has called "intellectual pollution" (Lakatos, 1970, fn. 1 on p. 176) if we would quit engaging in this feckless enterprise. I think that if psychologists would face up to the full impact of the above criticisms, something worthwhile would have been achieved in convincing them of it. Besides, before one can motivate many competent people to improve an unsatisfactory cognitive situation by some judicious mixture of more powerful testing strategies and criteria for setting aside complex substantive theory as "not presently testable," it is necessary to face the fact that the present state of affairs is unsatisfactory.

My experience has been that most graduate students, and many professors, engage in a mix of defense mechanisms (most predominantly, denial), so that they can proceed as they have in the past with a good scientific conscience. The usual response is to say, in effect, "Well, that Meehl is a clever fellow and he likes to philosophize, fine for him, it's a free country. But since we are doing all right with the good old tried and true methods of Fisherian statistics and null hypothesis testing, and since journal editors do not seem to have panicked over such thoughts, I will stick to the accepted practices of my trade union and leave Meehl's worries to the statisticians and philosophers." I cannot strongly fault a 45-year-old professor for adopting this mode of defense, even though I believe it to be intellectually dishonest, because I think that for most faculty in soft psychology the full acceptance of my line of thought would involve a painful realization that one has achieved some notoriety, tenure, economic security and the like by engaging, to speak bluntly, in a bunch of nothing. That is a bit much to expect that of anybody, even a psychology professor. In the case of graduate students, I find to my surprise a little more open mindedness on the point, although it can mean that a student has to change his doctoral dissertation topic from something that is more theoretically interesting to something less so but testable. It is my belief, after 45 years on the faculty at Minnesota, that well over half of the doctoral dissertations in soft psychology that are set up with the intention of testing an interesting causal theory are incapable of doing so. I don't see how any fair-minded person could dispute this who has sat on PhD final orals, even without having read my list of obfuscators!

However, despite my firm insistence that purely negative criticism of an intellectual boondoggle leading to Lakatos's "intellectual pollution" in the journals is an important form of academic husbandry, I do have some tentative suggestions for improving the situation. I am afraid that the best and clearest of them are still of a "negative" sort (e.g., critical editorial policies), but some of them offer a possibility of positive advance.

For investigators: Psychologists attempting to test a substantive theory in soft psychology should strive for a rationale by which an expected *amount* of effect could be predicted from the theory. Point values are ideal, but even in physics and astronomy they are surrounded by a tolerance based on an estimate of the experimental error. One hopes that, when enough persons become sufficiently skeptical about the weak corroboration provided by merely showing that the *X*s get higher scores than the *Y*s, that cheap and easy derivation might be replaced by one that says something about the range of non-null differences that would be consistent with the theory. At the very least, one might say that a theory accounting for less than such and such percent of the reliable variance is an uninteresting theory and does not deserve high priority for investigation, except for special considerations (e.g., a weak correlate of psychopathology that could serve as a genetic marker). I don't deny that there are cases in which small effects play a critical role in theory testing. But those are special cases in science. Because of the crud factor's ubiquity, merely saying that "there ought to be a difference between A and B" is a feeble test of anything, and we ought to work harder than we usually do to come up with some statement about points and ranges.

We should pay attention to Jacob Cohen's advice; given the bad effect of multiplying doubtful auxiliaries and *ceteris paribus* by the power function, I would push for higher statistical power than he did, perhaps saying that if you want to have a test of a theory you ought to set your sample size at a power of .9 or better. If there are two or more measures of a trait, the experimental or correlational design should include a discrepancy analysis in relation to interaction. I do not know whether a "standard" statistical method for doing this exists. Pilot studies ought to be fully reported. It should be emphasized in methodology courses that there is an ethical obligation, if one has done one or more pilot studies, *particularly pilot studies that were used to reject a possible line of investigation*, to publish all pilot studies that led the investigator to perform a large scale investigation. Even now, tradition requires that an unsuccessful attempt to replicate the main study should be reported, yet people do not always publish.

For editors, referees, journals: It would be helpful if journal editors regularly imposed the requirement of a successful replication, with certain exceptions such as studies that are terribly costly, or diseases that are very rare, or procedures that are dangerous. All statistical tables should be required to include means and standard deviations, rather than merely a *t*, *F*, or χ^2, or even worse only statistical significance. A table, offered for theoretical interpretation or for proposed clinical application of some device or procedure, that is confined to stating the significance level achieved and does not allow the reader to look at overlap, is as misleading and incomplete scientific reporting as failing to say from where you got your subjects, how they were chosen, or what their instructions were. I don't look upon this as a minor refinement that is merely pleasing to a perfectionist statistician. I look upon it as correcting a fundamental defect in our present habits (not true thirty or forty years ago in psychology) resulting from overemphasis of null hypothesis refutation. Confidence intervals for parameters ought regularly to be provided. If they cannot be, it should be said why not. In many circumstances it is possible to make a reasonable estimate of the percentage of variance accounted for by a given factor.

If the theory bears on some clinical problem such as a correlate or possible indirect indicator of a conjectured causal source for schizophrenia or whatever, appropriate alternative overlap statistics should be presented. A clinician who submits a paper advocating the use of a structured or projective test or some behavioral sample method for a given discrimination in psycho-pathology and does not offer an appropriate overlap measure, is unscholarly. For some purposes Tilton's (1937) overlap measure is all right, but as a clinician I would also impose the requirement — not a mere preference or suggestion but an absolute editorial requirement as part of complete scientific reporting — that the percent of one group reaching or exceeding the 10th, 50th, and 90th percentiles of the other group should routinely be reported.

It would be helpful to have a section of almost every journal reserved for publication of negative pilot studies. The shortest possible statement of the design compatible with scientific adequacy and an absolute minimum of theoretical discussion other than a brief statement of what motivated the pilot study being done, would add to people's yardage in a painless way and greatly increase the presently feeble and sometimes even negative motivation to publish negative pilot studies. This would also indirectly save a great deal of scientific time and taxpayer money. We must surely assume that many pilot studies which come out negative and hence might lead to abandonment of a once-promising line, probably have been done over and over again, especially by graduate students, because students do not know that some other investigator has already tried this and dropped it because he "failed to get an effect." One would have to make these papers short, easy, and painless, without too rigid criteria on quality, interest, or statistical power function.

For reviewers: Mention of statistical power should be obligatory in the review of every negative result. If it is objected that this is too much work for reviewers, then editors ought to adopt a policy of requiring that authors always state the statistical power. The present status of meta-analysis as a formalized method being still in dispute, I would not impose a requirement for it. But I think it fair to suggest that meta-analysis related to the auxiliaries ought to be helpful, even for readers who do not like the approach of Glass and Co. (Glass, McGaw, & Smith, 1981; see also Hunter, Schmidt, & Jackson, 1982). It helps us focus on the culprits, auxiliaries that might be responsible for giving a good theory a black eye. Reviewers ought to be sophisticated enough to know, and say explicitly in summary, that a mildly positive "box score" of tallies on reaching or failing to reach significance is *not* a strong sign of a theory's verisimilitude. The present reviewing practice is to do such a tally, after explaining away some of the positive and negative findings. Faced with the need to do some sort of integrated summary, the idea seems to be that, if a theory pans out with successful predictions appreciably more often than it fails, the box score speaks strongly in its favor. I cannot imagine any logician agreeing with this practice, as it fails to take into account the basic logical asymmetry between confirmation and falsification and pays no attention to the above list of obfuscators. Testing a theory in soft psychology in the light of those obfuscators and finding that its batting average is seven to three or six to four in the literature, while not totally worthless, is about as close to worthless as one can get for evaluating the theory's verisimilitude.

For theoreticians: I think we should be more optimistic about the possibility of making predictions beyond mere non-null difference predictions from rather weak theories. There are examples in the physical sciences in which at a given state of knowledge the theory was too weak or incomplete to permit derivation of numerical values, but was still capable of predicting rough function forms (see e.g., Eisberg, 1961, pages 49-51 on Wien's law). Sometimes what appear to be extremely weak general qualitative statements, incapable of generating anything numerical, turn out to generate quite interesting quantitative predictions when the applied mathematician goes to work on them, such as the relation of the sizes of certain second derivatives or regions in which there is a turn around or a flex point, or statements of that sort. Catastrophe theory is a recent example of astonishing quantitative richness. We should try harder for intermediate strength theories that, while they might not be capable of yielding point predictions, nevertheless yield statements about signs of derivatives, about inequalities without the parameters being known, about curve shapes, and so forth. It is, for example, sometimes possible to construct latent structural models of situations, as in my own current work in taxometrics, where the theory is far too weak to yield numerical values at the observational level but is still strong enough to yield a statement of numerical *equality* between two computed values based on observation (Meehl, 1973; Meehl & Golden, 1982).

One has the uneasy feeling that, if all this had been possible in soft psychology, it would have happened more than it has by now. While I cannot definitely refute that argument, I would emphasize that we do not work hard at doing something that is difficult, and different from our accustomed modes of thought, if we think that the way we are now doing it is working just fine! We are familiar with it, the other members of our troop of gregarious primates are busy doing it the same way, people get elected to high professional offices, and others receive various kinds of prizes for doing it this way. It is not surprising that clinical, social, and personality psychologists spend little time trying to figure out whether they could perhaps derive theorems about stronger consequences from semi-qualitative causal, compositional, or structural theories of the mind.

For teachers and doctoral programs: I think that PhDs in psychology should be required to learn a little undergraduate mathematics different from cookbook statistics. Inability to think mathematically among psychologists except in certain special areas is sometimes so gross as to be embarrassing to one familiar with the quantitative sophistication in other sciences. The Minnesota Department has been recommending mathematics courses to its undergraduate majors since I became chairman in 1951, with negligible results. Mathematics is hard, sociology is easy; we will never persuade the majority of psychology majors to take any mathematics unless we combine (a) a little mathematical content *used* and *on the final exams* in the courses they take with (b) explicit math requirements for our majors. Most arts colleges today offer undergraduate mathematics through calculus in a variety of forms, including some that are small in total hours required, and not so heavily geared to traditional problems of the physical sciences (like the volume of footballs) as when I was a student. There is an unfortunate circular feedback here at work. Since psychologists in the soft areas rarely know undergraduate mathematics, they do not think or talk mathematically as teachers, advisors, or research directors. As a result it is only natural that even a competent student forms the notion that for the kind of psychology he wants to do, knowing elementary mathematics is irrelevant. Busy people can hardly be expected to learn something that is a little difficult and quite time-consuming unless they can see its relation to what they intend to do; and they can't see any such relation if their mentors cannot do it because they never studied any mathematics either. I entertain the dismal conjecture that this is incurable, since my efforts to cure it (off and on over forty-five years) have had negligible local impact.

The question of what kind of mathematics psychology students should be thoroughly familiar with in connection with which research procedures, (as contrasted with having only a nodding acquaintance or being totally ignorant), is not easy to answer. But the common rationalization of mathematically ignorant psychologists ("Well, I understand the logic of factor analysis even though I don't understand the math") should not be tolerated in intellectually polite circles! The "logic" of a procedure like factor analysis is, of course, mathematical. There simply isn't any way you can "understand the logic" of the varimax solution to the rotation problem if you don't understand why there is a rotation problem. This is nothing but a rationalization by people who don't want to take the trouble to learn a little probability theory, vector algebra, or elementary calculus.

I am not merely being a purist about this. I have sat on PhD orals and read scientific articles by professors of renown that are fallacious in what they do with their quantitative results, because the theorist or investigator was so mathematically naive that it did not occur to him to ask, for instance, whether a certain function might be decelerated in a region and hence give rise to the appearance of an interaction effect, or whether his arbitrary choice of metric might determine the character of his results, or whatever. I would draw the line at requiring a psychologist who wants to use χ^2, for instance, to have fought his way through the proof of the theorems used in constructing the χ^2 tables. I don't think that fighting your way through all those gamma and beta functions (though one should know what a gamma function is!) sheds much of any light upon the properties of χ^2. Whereas if a student does not know that one of the more general ways of conceiving χ^2 is as a composite based upon summing the squares of variables that are themselves Gaussian, or if he doesn't know where that rule of thumb in statistics about "cells should have an expected frequency of 10 or more" comes from in terms of the underlying binomial construct, then he doesn't know what he ought to know as a scientist about his research methods.

I give you an extreme example which I think suffices to show that there is something the matter with psychology in this regard: Can anyone imagine a PhD in physics putting the Schroedinger equation on the blackboard, explaining how he was going to do his experiment in quantum mechanics, and then when asked what that funny little backward curlicue (like sort of a deformed lower case Greek delta) was, saying glibly "Oh, that's a partial derivative," then when asked what a partial derivative is, saying he didn't know? We don't even have to ask whether he would flunk the exam, because it is simply inconceivable to any informed person that a physics student would not know what a partial derivative is and does. You couldn't get a bachelor's degree in physics at West Overshoe Teacher's College if you didn't know that. Yet I have seen instances where a psychologist's doctoral dissertation consisted of factor analyzing somebody else's data, so that the student did not construct the test items, did not validate the test, did not test the subjects but found them in somebody's file, so that his sole intellectual contribution is that he did a factor analysis. He did it by using the varimax rotation, but when asked what the varimax rotation is can't even tell you who developed it, let alone what it does. You ask him whether he knows the relationship between John B. Carroll's breakthrough and Thurstone's simple structure criterion, and he hasn't a clue. You ask him how the use of fourth powers in this context is analogous to something Karl Pearson did about measuring leptokurtosis, and he hasn't the faintest idea what you're talking about. Now I think that to get a PhD by factor analyzing somebody else's test, administered by somebody else, so that your sole contribution is the factor analysis (via canned computer program) and providing a possible conceptual interpretation (usually very weak), when you do not understand the factor analysis, can best be described as scandalous.

It might help to require some reading of classic experiments and theoretical derivations in the other sciences, both biological and physical. I find that many psychologists literally do not know what a good theoretical derivation in a developed cumulative science looks like! While the mediocre students might not be grabbed much by this, I think that superior students will get the point and will become restive about the way in which soft psychology research goes about its business. It doesn't take many examples from chemistry, physics, genetics, physiology, and astronomy to bring a bright and intellectually alive student to the realization that these other scientists really have something in the way they get from the theory to the facts and back that is a lot more impressive — and, importantly for bright people, more intellectually satisfying — than the usual dismal prediction that the null hypothesis is false.

It would help if we could reduce the pathological emphasis on publication rate in regard to salary, tenure and promotion. One reason for the uncritical reliance on mere null hypothesis refutation as if it constituted a respectable test of the substantive theory is that it is a pretty safe way to spend one's time enroute to a publication. The change in the expectations of how much a student will have published already before his PhD between now and when I was in graduate school 45 years ago is frightening. The pressure is so great that I know students who are not intellectually dull or morally careless, who have sat in my office and said explicitly that while it was subject *X* that really interested them, they were putting in for a grant to study subject *Y* "because that's safer, and I'm sure to get grant support." I think this is pitiable and destructive. It is not only bad for the student's mental hygiene but in the long run it has a cancerous impact upon the discipline. But speaking either as a clinician or as an observer of the social scene, I am at a loss to suggest any remedy for it given the insane requirement today that nobody can be promoted or tenured in the academy unless he continues to grind out many papers [cf. the provocative and insightful book by Mahoney (1976)]. In evaluating faculty for raises, promotion, and tenure, perhaps there should be more emphasis on *Science Citation Index* counts, *Annual Review* mentions, and evaluation by top experts elsewhere, rather than on mere publication yardage. The distressing thing about this is that while academics regularly condemn "mere publication count," a week later in a faculty meeting or a Dean's advisory meeting they *are actually counting pages* in comparing Smith with Jones. This is a disease of the professional intellectual, resting upon a vast group delusional system concerning scholarly products, and I know my recommendations in this respect have a negligible chance of being taken or even listened to seriously. Since the null hypothesis refutation racket is "steady work" and has the merits of an automated research grinding device, scholars who are pardonably devoted to making more money and keeping their jobs so that they can pay off the mortgage and buy hamburgers for the wife and kids are unlikely to contemplate with equanimity a criticism that says that their whole procedure is scientifically feckless and that they should quit doing it and do something else. In the soft areas of psychology that might, in some cases, mean that they should quit the academy and make an honest living selling shoes, which people of bookish temperament naturally do not want to do.

Finally, I raise the delicate question — without pressing an answer, which I do not pretend to have — whether we should invest time and dollars in wide-ranging, large scale studies of the crud factor. Colleagues and students who have heard me lecture on the "ten malignant obfuscators" tend to focus on the crud factor magnitude as the weakest component of my argument. "The null hypothesis, taken literally, is always false in correlational (nonexperimental) studies" does not, of course, immediately imply "Pairwise correlations of arbitrarily (= atheoretically) chosen variables in most soft domains tend to run large enough to yield frequent pseudoconfirmations of unrelated substantive theories, given conventional levels of the statistical power function based on pilot studies." I daresay mathematical statisticians would look askance at the question "How big is the crud factor in Domain D?" given the unavoidable vagueness in specifying the variable set. What, one asks, is *the parameter* being estimated? "Estimating the crud factor" sounds too much like trying to find the tonic chord of the Universe.

However, given the sorry state of the art and the gravity of the problem, this attitude may be puristic. After all, we *do* commonly make similar rough-value statements in soft psychology. We allow ourselves to say such things as "SES usually correlates low to moderate with psychometric measures of ability and school achievement"; "Tests of so-called 'mechanical ability' correlate low to moderate and positive"; "Assortative mating coefficients vary from negligible to a high of around .50 and are almost never negative"; "Prediction of college grades (pre-inflation!) seems to have an upper limit, with a half-dozen predictors in the regression equation, of around .70." Such summary statements do not purport to be precise, but they are surely not empty of empirical content, nor are they useless. Since the proper baseline in testing substantive theories is not H_0 but the crud factor in a domain, so that a theory's "doing better than chance" is closer to "beating out the crud factor" than it is to "correlating nonzero," it is arguable that even a crude range estimate of the crud factor in a domain would be worth having. It requires very large *N*s and sizeable variable sets that are qualitatively heterogeneous and chosen with minimal theory in mind. Such data are hard to come by. How narrowly to specify "a domain" is a tough problem, and methods covariance within a domain (structured tests, projectives, interviews, ratings, ward behavior samples, work products, critical incidents, demographics, life-history facts) would surely not be equal. We might want to "bootstrap" the domain specification partly *post hoc*, in light of the mean and dispersion of pairwise correlations. It would be a lot of work and infected with much arbitrariness. Whether the resulting collection of crud factor values classified by domains and methods would be worth the trouble I am not prepared to say, although I lean slightly to "Yes" and will leave it at that.

All of these possible methods of improvement should be tried. In addition there is a more fundamental philosophical point to be raised, one which I have moral conflict about raising in my seminar every year and require reassurance about "tough love" from my colleague Lykken to get me to do it. *We should accustom ourselves and our students to the idea that there are some interesting causal theories in the soft areas that cannot presently be researched*, and it is arguably wrong to waste the taxpayer's money in state supported institutions to pretend to do it. I may mention briefly, without strong proof, a methodological issue in social science that deserves an article as long as this paper in its own right. There exists an implicit misconception, ubiquitous among students and professors studying soft areas, which could be cured or at least ameliorated by more extensive reading in the histories of the physical and biological sciences, together with a small dose of up-to-date philosophy of science. This misconception is that, if a theoretical conjecture is "scientifically meaningful" (not theological or metaphysical or so vague as to cover anything), then it must be possible to test it at the present time. Even a slight familiarity with the history of astronomy, physics, chemistry, medicine, and genetics shows that such a metatheoretical notion is plainly false. These other sciences are replete with examples of perfectly good "empirical" questions, askable by sophisticated scientists at a given time, that could not be answered given either deficiencies in the required auxiliary theories or the lack of an adequate instrumentation, whether for control of variables on the one side or, more commonly, measurement of variables on the other. A classical example of this for historians of science is August Comte's[8] description of the transition of human knowledge through the three phases of the theological or fictitious, the metaphysical or abstract, and the scientific or positive. Comte said that in the scientific or positive intellectual mode, it was obvious that there were certain things that human beings could never learn, such as the chemical constitution of the stars. For Comte, writing in the first part of the nineteenth century, the only way to find out the chemical constitution of any material body was to perform certain testing operations upon it in the laboratory such as exposure to reagents, litmus tests, and direct determination of precise weights and measures. Since one cannot put Alpha Centauri in a beam balance, or drop a chunk of it into a chemical retort, it seemed blindingly obvious to Comte that one could not ever find out its chemical constitution. It never occurred to him that the stars being hot gases that give off light, and the spectrum of light from an incandescent source indicating its chemical elements (spectroscopy had not been discovered in 1835), there could be an indirect way of determining stars' chemical compositions. He would have been stupefied to learn that this method is so precise that we know the percentage of various elements in the sun with a higher precision than we do for our own earth. A contemporary example from astronomy would be Feyerabend's suggestion about possible nonrelativistic departures from the Newtonian predictions of planetary motions, some of which might be explainable by a slightly altered estimate of the sun's oblateness, the measurement problem being that as the sun is not a smooth cue ball from a billiard table, we cannot by any instruments or methods available to us, or likely ever to become available, determine the oblateness of the sun to the accuracy that would be required. The most dramatic example from biological science in recent times, and one of the two or three greatest scientific discoveries ever made, is Crick and Watson's theory of the DNA. No amount of theoretical ingenuity would have enabled them to do this, let alone test it, until chemical methods were sufficiently precise to be able to show that in any organism the adenine and thymine are always precisely equal in the number of molecules present, as are the guanine and cytosine. It would not suffice to show that they are "correlated" or more-or-less equal. The important thing was demonstrating that the associated base pair were always precisely matched in number of molecules, to within a minuscule error of measurement. Nor would it have been possible to formulate such a theory in the first place until there was sufficient knowledge about the structure of these four organic bases, including

[8] Comte was the originator of "positivism" and is usually considered the founder of sociology.

exact details of the angles and distances between their component atoms, so that one could do the right kind of theoretical fiddling that Crick and Watson did for hours on end with their model pieces of cardboard and tin to see how they could fit together and have the right amount of stability. Finally, it was necessary that the technology of X-ray observation of extremely small physical systems should be so advanced that Wilkins could get pictures sufficiently detailed and clear to corroborate the conjectured helical structure. Even if somebody had by divine inspiration concocted the correct theory of DNA immediately after Thomas Hunt Morgan and collaborators presented the theory of the gene (forty years earlier), *there would have been no possibility of empirically testing it*. In physicists' current discussions of quarks and gluons, or astronomers' discussions about the Big Bang and about black holes, there are questions constantly being raised to which the answer is, "Yes, that's an interesting conjecture, but unfortunately we have no way of testing it at the present time, and perhaps we never will."

How do psychologists and sociologists come to be blind to this familiar fact about the more developed sciences, concerning the limitations on testing theories imposed by incompletely developed auxiliary theories and absence of measurement technologies? Some of my behaviorist friends consider it a fatal defect of Freud's theories of dreams or parapraxes that they cannot be presently tested in a rigorous quantitative manner, a claim with which I largely and cheerfully agree. I think that in order to test Freud's theories regarding the guiding of free associations during a psychoanalytic hour following the patient's presentation of the manifest content of a dream (see Meehl, 1970, 1983), we would probably require a more suitable type of statistical analysis than we presently have available, plus a well worked out and highly corroborated auxiliary theory of psycholinguistics, which we also do not have today. I think there are two reasons for mistakenly supposing that any "scientifically meaningful" or "truly empirical" theory must, if one is determined and ingenious, be strongly testable at the present time, over and above the optimism required to stay in the publishing business. First, there is the residue (or even the unmodified form) of 1929 operationism and logical positivism, which made the strange mistake of tying the very meaning or content of a scientific theory with the method of its verification. If one combines this notion of translation with insistence upon an available verification method and then further ignores the distinction that even the Vienna Positivists made between logical unverifiability, empirical unverifiability, and technical unverifiability at a given point in time, it seems to follow that if a theory is "meaningful" (i.e., not metaphysical or theological or tautological) by positivist standards, then it is *ipso facto* testable.

Suppose one abandons the notion of complete conceptual reducibility of all concepts to observable predicates or functors, hence sentence verifiability as a meaning criterion, and hence operationism as a requirement for theoretical definitions. A proper subset of theoretical concepts may be operationally defined, but the great majority of them are not so, being defined only contextually by the mathematical network, if such exists, along with the interpretative text, that text *not* being confined to ostensive linkages. Even if one accepts the last ditch positivist effort at defining an empirical meaning criterion (Carnap, 1956), it says that a theoretical term is meaningful because it appears in at least one derivation chain somewhere. This gives rise to a *concept empiricism* upon which a *statement empiricism* is erected. Roughly, this says that, if a theoretical statement is well formed syntactically from such meaningful concepts (and perhaps certain further metaconstraints semantically?), then the statement is scientifically meaningful, even if it does not itself appear in a derivation chain terminating in a pure observational statement and, hence, is neither confirmable or falsifiable. If cross-eyedness is recessive in the Siamese cat (I have not looked this one up but its correctness doesn't matter in our example), and a neutered stray cat of unknown lineage comes to live with us the question "Does he carry the recessive gene for cross-eyes?" is a perfectly legitimate scientifically meaningful query, although we have no means of answering it.

Now if a pre-1956 statement empiricism is subscribed to uncritically, as it is by many psychologists raised in an outdated philosophy of science, a person so mal-instructed can then start with the old notion that, if a sentence is scientifically meaningful, it must be completely reducible to observational statements. He then connects these observational statements with some counter null statistical hypothesis, so that he thinks that proving the one proves the other. It follows that any meaningful theoretical statement one can make must be testable by H_0-refutation on appropriate data. Whereas the truth of the matter is that many meaningful theories, including theories that most of us would consider quite interesting intellectually and of great theoretical importance to find out about, simply cannot be tested at the present time, either because they are embedded in a vast net of highly problematic auxiliaries and *ceteris paribus* clauses or because we have no adequate technology of measurement. Sometimes a theory is untestable because the technology is too loose (as in the obfuscator about detached validation claim above), or sometimes it's because there is literally no measure available.

I frequently have the experience where a student asks me to serve on a doctoral examining committee, tells me about a design aimed at testing a theory in soft psychology, and my heart sinks as I listen. A great cloud of cognitive gloom descends upon me, because the thought that keeps coming into my mind is "You can't test it like that, you'll just never manage to test it like that." But if I try to explain to the student why it can't be tested, he takes me to mean that it somehow is an illegitimate theory, or a theory that is "metaphysical" or permanently beyond our ken, which is not my claim at all. The problem is that the majority of theories in soft psychology are related to the data in somewhat the same way as the constitution of the stars was to the data extant before anybody discovered spectroscopy in the terrestrial lab. *Point: We should maturely and sophisticatedly accept the fact that some perfectly legitimate "empirical" scientific theories may not be strongly testable at a given time, and that it is neither good scientific strategy nor a legitimate use of the taxpayer's dollar to pretend otherwise*.

Addendum

Local readers of this manuscript have asked why I do not discuss meta-analysis (Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982), and I would not want to leave the impression that I am unaware of it or view it unfavorably. On the contrary, I think meta-analysis is one of the most important methodological contributions of this generation of psychologists, arguably *the* most important, and have so stated to Professor Glass in our correspondence. This paper could be viewed as helping make the meta-analysts' case against the conventional narrative, impressionistic, "box-score" approach to reviewing research. However, the chief reasons for my not discussing meta-analysis were these:

1. Meta-analysis has become a highly technical ramified system of conceptual and mathematical issues, many articles and several whole books being devoted to it. To discuss these matters briefly and superficially would be inappropriate. To do it in depth is precluded by space limits on an already longish piece, as well as being beyond my statistical competence.

2. Meta-analysis was developed to study outcomes of interventions (e.g., influence of class size, efficacy of psychotherapy or psychotropic drugs) rather than as a method of appraising the verisimilitude of substantive theories. We do not normally assume *theoretical corroboration* to be a monotone function, even stochastically, of *effect size*; and in developed sciences an observed value can, of course, be "too large" as often as "too small."

3. A representative ("typical") effect size, whether of aggregated or disaggregated studies, is interpreted or qualified in meta-analysis via estimates of its standard error, emphasizing its trustworthiness as a numerical value. This statistical stability (under the laws of chance) is a very different question from how closely the effect approximates a theoretically predicted value. More importantly, it does not ask how "risky" the latter was in terms of the theoretically tolerated interval, in relation to the *a priori* range of possibilities. These two questions, taken jointly as the basis of all theoretical appraisal, require a different approach from that of evaluating technological outcomes in a pragmatic context.

Whether a quantitative index of the "corroborative increment" given a theory by a particular experiment can be constructed is doubtful, but I am currently exploring that notion (Meehl, 1990). Briefly: A theory predicts an interval, *I* (= intrinsic tolerance) for a numerical value observable in a specified experimental setup. (The intrinsic tolerance could be widened by the standard error, or by 2 *SE* [based on the data], to yield an adjusted tolerance, but I do not favor that adjustment.) The ratio of this to the *a priori* range of values *S* (= *Spielraum*) is termed the relative tolerance, *I/S*; and the latter's complement $(1 - I/S)$ is the theory's intolerance, *In* (i.e., the Popperian experimental prediction "risk"). The deviation *D* of the observed value x_o from the edge of the tolerated interval is also divided by *S* to give the relative accuracy, *D/S*. The corroborative increment associated with this experiment is defined as $C_i = (1 - D/S)(1 - I/S)$. Over *m* studies, the statistics (m, M_C, σ_C) jointly provide a basis for appraising the theory, and disaggregating with respect to fact domains should provide leads for modifying the theory, auxiliaries, or *ceteris paribus* clauses.

REFERENCES

BAKAN, D. (1966) The test of significance in psychological research. *Psychological Bulletin*, 66, 423-437.

BRUNSWIK, E. (1947) *Systematic and representative design of psychological experiments*. (University of California Syllabus Series, No. 304) Berkeley, CA: University of California Press.

CAMPBELL, D. T., & FISKE, D. W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

CARNAP, R. (1956) The methodological character of theoretical concepts. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science: I. The foundations of science and the concepts of psychology and psychoanalysis*. Minneapolis, MN: University of Minnesota Press. Pp. 38-76.

COHEN, J. (1962) The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal and Social Psychology*, 65, 145-153.

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

EISBERG, R. M. (1961) *Fundamentals of modern physics*. New York: Wiley.

GLASS, G. V., MCGAW, B., & SMITH, M. L. (1981) *Meta-analysis in social research*. Beverly Hills, CA: Sage.

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

HULL, C. L., HOVLAND, C. I., ROSS, R. T, HALL, M., PERKINS, D. T, & FITCH, E G. (1940) *Mathematico-deductive theory of rote learning*. New Haven, CT: Yale University Press.

HUNTER, J. E., SCHMIDT, F. L., & JACKSON, G. B. (1982) *Meta-analysis: cumulating research findings across studies*. (*Studying organizations: innovations in methodology*. Vol. 4). Beverly Hills, CA: Sage. [ref. amended]

LAKATOS, I. (1970) Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge, Eng.: Cambridge University Press. Pp. 91-195.

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

LYKKEN, D. T. (1968) Statistical significance in psychological research. *Psychological Bulletin*, 70, 151-159. [Reprinted in D. E. Morrison & R. E. Henkel (Eds.), *The significance test controversy*. Chicago, IL: Aldine, 1970. Pp. 267-279.]

MAHONEY, M. J. (1976) *Scientist as subject: the psychological imperative*. Cambridge, MA: Bollinger.

MEEHL, P. E. (1967) Theory-testing in psychology and physics: a methodological paradox. *Philosophy of Science*, 34, 103-115. [Reprinted in D. E. Morrison & R. E. Henkel (Eds.), *The significance test controversy*. Chicago: Aldine, 1970. Pp. 252-266.]

MEEHL, P. E. (1970) Some methodological reflections on the difficulties of psychoanalytic research. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science: IV. Analyses of theories and methods of physics and psychology*. Minneapolis, MN: University of Minnesota Press. Pp. 403-416. [Reprinted *Psychological Issues*, 1973, 8, 104-115.]

MEEHL, P. E. (1973) MAXCOV-HITMAX: A taxonomic search method for loose genetic syndromes. In P. E. Meehl, *Psychodiagnosis: selected papers*. Minneapolis, MN: University of Minnesota Press. Pp. 200-224.

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

MEEHL, P. E. (1983) Subjectivity in psychoanalytic inference: the nagging persistence of Wilhelm Fliess's Achensee question. In J. Earman (Ed.), *Minnesota studies in the philosophy of science: X. Testing scientific theories*. Minneapolis, MN: University of Minnesota Press. Pp. 349-411.

MEEHL, P. E. (1990) Appraising and amending theories: the strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry*, 1, 108-141, 173-180.[ref. updated]

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

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

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

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

ROSENTHAL, R. (1966) *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.

ROZEBOOM, W. W. (1960) The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57, 416-428. [Reprinted in D. E. Morrison & R. E. Henkel (Eds.), *The significance test controversy*. Chicago, IL: Aldine., 1970. Pp. 216-230.]

SCHILPP, P. A. (ED.) (1974) *The philosophy of Karl Popper*. LaSalle, IL: Open Court.

SWENSON, W. M., PEARSON, J. S., & OSBORNE, D. (1973) *An MMPI source book: basic item, scale, and pattern data on 50,000 medical patients*. Minneapolis, MN: University of Minnesota Press.

TILTON, J. W. (1937) The measurement of overlapping. *Journal of Educational Psychology*, 28, 656-662.

Accepted January 23, 1990.

pdf by ljy January 2003