In D. W. Fiske and R. A. Shweder (Eds.) *Metatheory in social science: Pluralisms and subjectivities*. Chicago: University of Chicago Press, 1986. (Chapter 14, pp.315-338)

What Social Scientists Don't Understand

P. E. Meehl

In the papers prepared for the conference on which this book is based and in the discussion, there were some matters almost universally agreed upon but repeated unnecessarily. Then there were some things that should have been agreed upon but were not. Finally, there were matters that were not agreed upon that needed more intensive examination—matters playing a central role in the philosophy of the social sciences and in the present intellectual gloominess that seems to prevail in all of the disciplines represented.

THINGS GENERALLY AGREED ON FROM THE BEGINNING

It was agreed that logical positivism and strict operationism won't wash. Logical positivism, in anything like the sense of Vienna in the late twenties turned out not to be logically defensible, or even rigorously formulatable, by its adherents. It is epistemologically unsound from a variety of viewpoints (including ordinary-language analysis), it is not an accurate picture of the structure of advanced sciences such as physics, and it is grossly inadequate as a reconstruction of empirical history of science. So it is dead. All old surviving logical positivists agree, including my friend and teacher Feigl, who invented the phrase "logical positivism" and was the first to introduce the approach in the United States in 1931. The last remaining defender of anything like logical positivism was Gustav Bergmann, who ceased to do so by the late 1940s.

Why, then, the continued attack on logical positivism and its American behaviorist near-synonym "operationalism"? My answer to this is unsettling but, I think, correct. Our conference on social science came about partly because of widespread dissatisfaction about the state of the art, and we have always been more introspective methodologically than the physicists and biologists, who engage in this kind of thing only under revolutionary circumstances. My perhaps cynical diagnosis is that one reason the conference members spent needless time repeating that logical positivism and simplistic operationalism are incorrect views of scientific knowledge is that this relieves scientific guilt feelings or inferiority feelings about our disciplines. It's as if somebody said, "Well, maybe clinical psychology isn't up to the standards of historical geology or medical genetics, let alone theoretical physics; but we needn't be so fussy about our concepts and empirical support for them because logical positivism, which was so stringent on that score, is a nefarious doctrine and we are no longer bound by it."

I see a connection here with the concern about the social sciences as "science" and what it takes to be really scientific. I have never had much interest in this labeling question. I don't see how anybody familiar with various disciplines in the life sciences, and even some of the "inorganic" sciences, could see it as a clear-cut or interesting question. Skinner points out in *The Behavior of Organisms* (1938, pp. 41,419) that the curves obtained from a single organism in the operant conditioning chamber ("Skinner box") are smoother and more reproducible than many of the curves obtained by medical students in their introductory physiology lab course. The verbal report of a sophomore, experiencing for the first time a negative afterimage, is reproducible enough so that you can afford to bet $10,000 at 100-to-l odds that a subject pretested for having normal color vision and not insane will report that what he sees after presentation of a red circle and being asked to fixate a distant gray wall is a large, faded blue-green circle. Avoiding the mass media identification of the word "science" with test tubes, and instead attending to reproducibility, degree of quantification, and conceptual neatness (while allowing open concepts), my own discipline of psychology has subareas that are highly scientific, such as some branches of visual perception, cognitive processes, behavior genetics, and animal learning. Others are of intermediate "scientificalness," like trait theory, differential psychology, and psychophysiology of emotion. Still others, like projective techniques and psychoanalytic dream interpretation (which I practice), do not deserve to be called scientific at all. Some branches of historical geology, especially the paleontology involved in evolutionary theory, are much overrated as to their scientific status, being in a primitive degree of quantification, with highly speculative components. That doesn't mean we call geology or paleontology unscientific. Some branches of organic medicine are in a primitive state, while others are almost in as advanced a scientific state as chemistry or physics. It seems to me pointless to argue about these matters and mentally unhealthy for a social scientist to get involved in the semantic hassle as to whether he is engaged in science or not. It would be desirable to strike that question (perhaps even the very term "science") from the methodological vocabulary of conferences like ours. In addition to being hygienic, this would conduce to greater intellectual honesty because it would force us, instead of warding off semantic attacks by administrators or skeptical intellectual laymen with an antiscience bias, simply to ask the question "To what extent does this discipline contain knowledge that brings some sort of credentials with it?" Whether there is any kind of credentialed knowledge that is not, in some carefully specified sense, "scientific" knowledge is an interesting philosophical query, but one we need not answer for our purposes.

In psychology, at least, there is a tendency to conflate "experimental" with "quantitative" and then in turn to mix up "statistical" with "mathematical" as if they were the same thing, which they are not. Social science has many more statistical methods and findings than it does mathematical models, except in economics, where the connection between the mathematical models and the empirical statistics is regrettably tenuous. One can proceed experimentally without being quantitative; one can be quantitative in the sense of statistical counts without having a mathematical model for the causal structure underlying the processes; one could be highly quantitative, both in the sense of statistics of observations and in the sense of the postulated causal model being formalized in differential equations, and not be capable of experimenting at all, as in astronomy. Even worse is the conflation of the word "empirical" to mean "experimental and quantitative," a tendentious mistake that begins by failing to look up the word in Webster (for more detailed discussion of this, see Meehl 1983).

Another thing that seemed universally agreed upon is that a classical Newtonian model of science is incorrect. In my college days, nobody thought that Hull (1943) or Skinner (1938) or Thurstone (1935) was "the Newton of psychology," or that there was even ever going to be such a person. There are at least three broad classes of scientific theories, and all three kinds are to be found in all three major divisions of science (physical, biological, or social). First, there are *functional-dynamic theories*, Newton's, the theory of heat, Skinner's laws of learning, or the mathematical principles of population genetics being examples. Many of these, and certainly the impressive ones, are expressed in mathematical form and look like the Newtonian business, but not always. Functional-dynamic theories relate states to states or events to events and are most like Aristotle's concept of efficient causes. We say that when one variable changes, certain other variables change in such and such ways; their ideal form is the partial differential equation. Functional-dynamic theories do not, however, have to be completely general in the way that classical mechanics or thermodynamics purport to be. If, for instance, Hull's 1943 book had held up, we would have been overjoyed, even if those equations for habit strength held only for mammals and not for fishes or invertebrates. No one would have been disappointed by such departures from generality. What was disappointing is that the basic qualitative aspects (e.g., even the necessity for reinforcement) did not hold up—the latent learning experiments falsified it. It was by no means clear that the function form (simple positive growth function) was correct.

What scientists hoped (and believed?) was that the qualitative listing of intervening variables in Hull's famous set was general over at least mammalian species and that these variables would all be growth or decay functions of input like number of reinforcements, or extinction trials, deprivation time, and so on. What turns out from the research, especially that of the Skinnerians, is that no such mathematical construct as habit strength can be justified, even within a single species, the white rat, and in the controlled conditions of the laboratory on which the theory was based. So the distressing thing about Hull's heroic effort was not that it doesn't work for earthworms or that sometimes it's hard to estimate the growth constant or that for some species the reinforcers are strange. The disappointing thing about Hullian theory was that it didn't even hold up in the Skinner box or in the maze for the white rat, given hunger as a drive and food as a reinforcer. That is, on the very species, drives, rewards, and historical variables that were its origin, the theory did not do an adequate job, even as first-level description, let alone as causal analysis.

Second, there are theories that are *structural-compositional*. Their main idea is to explain what something is composed of, or what kind of parts it has and how they are put together. I think it is unfortunate to speak, as some did at the meeting or in papers, about reductionism as an evil force, to be exorcised by incantation. Some of the most impressive achievements in other sciences have been highly reductionist, and there are disciplines in which the reductionist aim could be almost described as the main scientific program (e.g., biochemical genetics). The main thing Crick and Watson made possible with the DNA business (in itself, mainly a big piece of reduction) was an entire research program, a proliferating enterprise of Kuhnian "normal science," in which the whole point is to reduce. What biochemical geneticists are mainly engaged in, other than the technology of manipulation, is precisely the reduction of something called a "gene," previously understood solely in terms of the mathematics of linkage and population genetics plus approximate location on a chromosome, to a specified sequence of codons. Whether or not the phenomena of a domain can be successfully reduced to those of another is an empirical question.

I suspect part of the problem here is our fear, connected with trade union considerations or vested interests, that if somebody, somewhere, somehow, someday were to reduce one's favorite concepts to lower-order ones, one's enterprise would have been shown feckless, which is of course absurd. Even when an almost complete job of reduction has been carried out in a highly corroborated theory in the physical sciences, nobody in his right mind proposes to liquidate discourse carried out for certain purposes at a higher level. We know what goes on about heating a building in terms of kinetic theory, but when a heating engineer talks with an architect or economist, he does not formulate the subject matter in terms of kinetic energy of molecules. I have always thought it foolish for some of my colleagues in sociology to get upset at the possibility that some psychologist might claim that all social symbolic interactions must ultimately be "based upon psychology," that is, reduced to individual organismic principles of learning (Skinnerian or otherwise). I think this kind of worry is partly what seduced such a smart and methodologically aware scholar as Durkheim into making some elementary bloopers in statistical inference in his classic book *Suicide*.

People who have thought through the "pyramid of sciences" problem are free of the worry—witness Skinner. As a good solid materialist and biological psychologist, Skinner entertained not the faintest doubt that every one of his learning principles involved structural changes in CNS. When pressed in conversation, he could argue, "When the laws of behavior are sufficiently worked out in mathematical detail (in the next generation following me), and when the anatomy, especially the microanatomy, and the physiology of the brain are very thoroughly understood, there will be no problem of prematurely forcing a speculative translation because it will be perfectly apparent how the brain/behavior dictionary will read." Similarly, Freud, who through his entire life always assumed that basically everything happening was in the brain and attempted an ambitious project along those lines, was perfectly clear in his arguments as to why brain language was not useful at this stage of our knowledge of psychopathology. The appropriate language in which to discuss the mental machinery is mental.

Third, some theories are *evolutionary* (historical or developmental), for example, Darwin's theory, Wegener's theory of continental drift, Freud's early theory of the life-historical origins of the obsessional versus the hysterical neurosis, historians' theories as to the fall of Rome.

THINGS THAT SHOULD HAVE BEEN AGREED UPON

The question as to whether, how much, and what kind of quantification is useful is again an empirical one not to be decided from the philosophical armchair, nor on the basis of either a scientistic obsession with mathematics or worship of physics on the one side or a humanist antiscientist bias on the other. The question is whether or not the use of a certain theoretical formalism, or the use of a certain kind of statistical procedure in data reduction, does or does not help matters. It is necessary to think clearly about words and to realize that many of the words—I would say most words—both in ordinary language and in scholarly discourse that purport to explain anything are quantity words intrinsically. Not always, but almost always. That Freud doesn't express libidinal cathexis in *lib* units doesn't detract one whit from the fact that part of his explanatory system involves making comparisons between quantities of it. When a social scientist speaks of something—anything, a tribal custom or suicide tendencies or unconscious memories or a white rat's lever-pressing disposition—he typically uses words like "always," "frequently," "typically," "rarely," "never," "oddly," "weakly," "under special conditions," "mostly." Every single one of these words is a claim of the degree to which some force or entity exists or influences; every single one indicates a frequency or probability with which something happens or the magnitude of a disposition (propensity). It is foolish for social scientists to try to get away from this simple fact about the descriptive language of their disciplines. The question is, What are the circumstances under which it pays off at a given state of knowledge to re-express these quantity words of ordinary English in explicitly numerical form? How strong is the claim that can be made for the resulting metric—that is, does it have certain nice properties such as a ratio scale or interval scale? Does it demand them?

There should have been more agreement on the legitimacy of open concepts. The writings of Waismann (1945), Pap (1953; 1958, chap. 11), and Carnap (1936, 1937) should have taken care of that question. The famous "testability and meaning" paper, whatever Carnap's main intentions at the time, was at least partly a logician's explication and justification for the use of open concepts in science, and Carnap subsequently became much less operational than he was in that paper. I had thought that a sufficiently satisfactory exposition of it for many social science disciplines was provided by Cronbach and myself thirty years ago (1955).

As Pap and others have clearly shown, considerable openness of concepts exists even in the most advanced sciences, and a great deal in the primitive early stages of any science; they abound in the life sciences. There isn't any good reason for saying that social scientists may not employ them once you have seen that the problems are methodological. My own view is that the proper way to deal with open concepts in psychology, including partially defined constructs like the heritable component of *g*, lies in the application of an appropriate stochastic mathematics. We don't get rid of open concepts by pseudooperational definitions, nor ought we to rejoice (like an obscurantist) in their persisting openness. What we ought to do is to tighten the stochastic nomological net increasing the strands, improving the instrumentation, and thereby reducing the stochastic feature. There is no reason to require that the end state of such process be to liquidate probability notions from the object language, which is fortunate, because there is no hope of doing so. There is still a little bit of confusion between probability as an epistemic concept, referring to the degree of evidentiary support or corroboration for some fact or theory, and probability as a metrical concept of the object language, built into the theory itself and considered likely to remain (regardless of the future state of evidence) because it is part of the theoretical substance. This second meaning of "probability" is not epistemological but has either a purely formal frequency interpretation or, as I would prefer, a Popperian propensity interpretation.

Should anyone still worry about the legitimacy of a contextual or implicit definition of an inferred causal entity based on covariation of observations? I had thought that was settled thirty or forty years ago, when philosophers of science (partly the positivists themselves and certainly their critics and amenders) recognized that only a proper subset of theoretical terms is directly "coordinated" to observational functors and predicates. The rest of them, the majority, acquire their meaning via a complicated mixture of at least three components: (a) their formal role in the theoretical network, a contextual partial interpretation similar to the implicit definitions of point and line in geometry; (b) the fact that the net as a whole is tied to minimally theoretical or pure observational terms, called "upward seepage"; and (c) informal explications in an associated text that uses models, analogies, even occasional mentalistic metaphors to contribute to meaning.

As to realism, I have never met any scientist who, when doing science, held to a phenomenalist or idealist view; and I cannot force myself to take a nonrealist view seriously even when I work at it. So I begin with the presupposition that the external world is really there, there is a difference between the world and my view of it, and the business of science is to get my view in harmony with the way the world really is to the extent that is possible. There is no reason for us to have a phobia about the word "truth." The idea that you shouldn't ask whether a scientific statement is true, separate from the anthropologist's or the Hogo Bogos' belief in it, because you can't be absolutely certain, is a dumb argument, refuted by Carnap in his famous replies to Kaufmann (Carnap 1946a, 1946b, 1948; see also Carnap 1949). Nor does it denigrate any culture's (or subculture's) values or forms of life to say that its denizens are mistaken as to certain causal facts. The point is not to conflate our admiration for the Hogo Bogo form of life or Faustian man's techniques in the moon shot with a factual question.

Regarding the imputation of motives or intentions on the basis of fallible behavior indicators, I believe it correct to say that the basic ideas involved were, for psychology, handled pretty well by Tolman (1932). In fact, Tolman relied heavily, although with a greater emphasis on the preeminence of the "docility" criterion, upon a powerful analysis by McDougall (1923) sixty years before our conference (cf. Murray 1938, pp. 54-76). So I found it a little discouraging that social scientists were reinventing McDougall's wheel after six decades. If the McDougall or Tolman kind of analysis (spontaneity, persistence with variation, cessation on goal reaching, anticipatory preparation for goal situation, improved performance contingent on getting results [= docility]) strikes some as too mentalistic (an objection I myself would have trouble understanding, let alone accepting), one could turn instead to Skinner (1938), whose discussion of why we introduce state variables like drive and emotion after doing our preliminary analysis of behavior in terms of reinforcement and extinction is highly sophisticated and adequate to most—I do not say all—scientific purposes. Or for that matter, one may consult Freud's 1916 paper on the unconscious, where his discussion of how the rationale of imputing unconscious motives to oneself is at bottom no different from that of attributing motives, conscious or unconscious, to other people.

I am not such an optimist as to suppose that there are no technical problems involved here. One could say that a souped-up generalization of the Campbell-Fiske multitrait-multimethod matrix is the answer, with plenty of room for time factors and cross-lag correlations included in the statistical tool kit. But my main point here is that the subject was discussed at the conference as if nobody had ever thought of it before. Furthermore, there was confusion between the problem of vagueness of meaning (if you like, "partial interpretation" or "open concept" definition) where a motive is imputed and the probabilistic character of the inference to it from fallible behavioral indicators. A motive is itself an open concept, and it is partly because of that openness—that is, the stochastic character of the strands in the nomological net and even the qualitative incompleteness of the net itself—that there is epistemic doubt, in some cases very great.

Of course, it is silly to think that the fact that you can have warranted doubt about an empirical statement somehow makes it illegitimate, and even sillier to think that there should be differences in degree of doubtfulness under such circumstances. I don't think that Tolman or McDougall or Skinner or Freud needed to be philosophically sophisticated or to use any technical philosophy-of-science jargon in expounding this, but it is perhaps a help in this day and age to have some of that available. So here we have another example of where a rather unphilosophical, tough-scientist approach will enable you to make progress, as will a highly sophisticated and technically competent philosophy-of-science approach.

Nobody today holds a strict deductive model, including Hempel, who propounded it. However, nobody has succeeded in presenting a useful meaning of scientific explanation that is totally unlike the Hempel model. All explanations that people are willing to take seriously look somewhat like that model, given the allowance of statistical laws together with suitable *ceteris paribus* clauses. Until somebody shows us what an explanation is that differs radically from a modified deductive model, I am not going to be impressed with the admittedly valid criticism of Hempel. This is made easy for me—and, I gather, difficult for some people—

because I take it as perfectly appropriate to overlap "causes" with "reasons," and I totally reject the ordinary-language claim that this cannot be done. Abstract reasons in Plato's heaven are not causes. But the hearing of reasons, stating of reasons, believing in reasons, tokening of sentences that mean reasons are events in the world and partake in the causal order. I would maintain with Tolman that the difference between a lawyer making a complicated argument to the Supreme Court in order to protect the taxpayer and a rat turning right in order to get food at the end in the goal box has not been shown to be other than a difference of very great and impressive degree, that is, not of kind. The imputation of motive to a person performing an instrumental act does not differ in any essential qualitative way from the implication of an unconscious motive, as in psychoanalysis, or a nonworded motive to the white rat. It's a matter of vagueness and a matter of degree of evidentiary support.

TESTING A WEAK THEORY

It seemed to me that our preoccupation with the demise of positivism and our worries (or self-reassurances) about obtaining some sort of "scientific ideal" in the social sciences led us to spend considerable time on the wrong things rather than on the really important topics. The latter are mostly variations on the same theme, namely, how weak theories can be strongly tested. I agree with Sir Karl Popper that talking about the meanings of words, or more sophisticatedly, about the nature of scientific concepts, is almost always a waste of time.

The important thing to clarify is the structure of the theoretical network and the resulting empirical tests. When do reductionist strategies work? How does one tell whether they are working, preferably early on, before a lot of time is wasted on them prematurely? Is it possible to concoct fairly strong tests of weak theories, and how should this be done? When is quantification worthwhile and when is it premature, or even fake? Can there be a general strategy for sequencing research studies of interaction effects of high order with an eye to generalizability? How should we relate the power function in statistical significance testing to the desire for strong falsifiability, since we know that the null hypothesis in social science is always false? Is meta-analysis a satisfactory solution of the crude pro/con count of studies found in a typical *Psychological Bulletin* review article? Is Lakatos's distinction between progressive and degenerating research programs, despite recent criticisms of it, a worthwhile distinction that could be helpful to the social scientist? How should the *ceteris paribus* clause be used in early stages of theory building? In saving what might be a good theory from premature death by quick falsification, when we use what we hope is "legitimate ad hoc-ery," can any useful rules of thumb be stated about degrees and kinds of ad hoc-ery, such as Lakatos's three kinds? Are any generalizations possible from the history of science about the fruitfulness or wickedness of ad hoc-ery of various sorts at different stages in the testing of a theory?

Obviously each of these topics deserves a separate paper longer than the present one. So I confine myself to a more intensive discussion of only one, the biggest one, as I see it. I shall propound a controversial thesis, deliberately stated in strong language.

> Thesis: Owing to the abusive reliance upon significance testing—rather than point or interval estimation, curve shape, or ordination—in the social sciences, the usual article summarizing the state of the evidence on a theory (such as appears in the *Psychological Bulletin*) is nearly useless.

The distribution of obtained significant and nonsignificant results is an arbitrary and complex artifact of eight methodological factors largely unrelated to a theory's verisimilitude, namely, (a) experimental design, (b) inherent construct validity of measures, (c) reliability of measures, (d) properties of the statistical power functions, (e) presence and size of higher-order interactions, (f) verisimilitude of auxiliary theories relied on in deriving empirical predictions, (g) differential submission rate of manuscripts reporting significant versus nonsignificant findings, and (h) editorial bias as to the same. The net result of these influences on the pro/con count is that usually such a heap of studies is well nigh uninterpretable.

Colleagues think I exaggerate in putting it this way. That's because they can't stand to face the scary implications of the thesis taken literally, which is how I mean it. Even though it is stated in all good elementary statistics texts, including the excellent and most widely used one by Hays (1973, 415-17), it still does not seem generally recognized that the null hypothesis in the life sciences is almost always false—if taken literally—in designs that involve any sort of self-selection or correlations found in the organisms as they come, that is, where perfect randomization of treatments by the experimenter does not exhaust the manipulations. Hence even "experimental" (rather than statistical or file data) research will exhibit this if interaction effects involving attributes of the persons are studied. Consequently, whether or not the null hypothesis is rejected is simply and solely a function of statistical power.

Now this is a mathematical point; it does not hinge upon your preferences in philosophy of science or your belief in this or that kind of theory or instrument. Whether investigator Jones, in testing theory *T* with a predicted observational relationship, succeeds in refuting H_0 depends upon the eight factors listed above. Expanding a bit on these, the region of the independent variable hyperspace in which the levels of a factor are chosen is something Fisher didn't have to worry much about in agronomy, for obvious reasons; but most psychologists have not paid enough attention to Brunswik on representative design.

There will usually be wide variation over studies in the intrinsic qualitative construct validity of the measures (both of input and output variables). The reliability of the measures will typically vary widely from values as low as .40 to as high as .95, and hence highly variable upper bounds are set on the net construct validity after attenuation by unreliability. As a result, the ordering of two measures as to their net attenuated construct validity may be quite different from the ordering as to their intrinsic qualitative construct validity because of marked differences in reliability.

While sample size exerts the biggest impact upon a statistical power for a given degree of real difference, *N* will be only partly a function of rational considerations stemming from the research problems but heavily a function of historical and geographical accidents, chronological age and status of the investigator, whether a study is a Ph.D. dissertation or part of a five-year project. Now the frightening (and hence repressed) point is that when we scramble these different factors and get a net power function for refuting H_0, the relationship between the probability of successfully refuting it and the verisimilitude of the substantive theory can hardly be large. I know of no realistic way of saying how small it could be, but any such relation would necessarily be markedly attenuated because of these other factors, none of which is of small magnitude in its impact.

This effect would be there even if the size of a difference in standard score form were itself a monotone function of the verisimilitude of the theory, which nobody claims it is. Most psychological theories in the "soft areas" of psychology do not even attempt to say how large an effect ought to be if perfectly measured, or even whether the theory implies that the main effect of a theoretical variable should be bigger than that from other compatible theories that contribute to determining what happens in the domain. For example, I might think that failure would have a different kind of effect on upper- and lower-class teenagers but that there would also be an interaction with IQ and also with the kind of task in which they failed (which already gives me some bad interaction problems). But even if I thought I had a powerful theory, I would not be likely to say that the self-concept from social class is therefore the biggest factor influencing a child's response to failure or success. And I certainly would not be in a position to say anything metric about these values. Even if the hapless reviewer is more sophisticated than most of them seem to be, when he looks at this pattern of "$p < .05$," and "$p < .01$," and "*n.s.*," he is not in a position to judge the extent to which the obtaining or nonobtaining of a statistically significant effect is artifactual with respect to the testing of the theoretical substance. So it seems to me that the first step is to get that message across to psychologists who write and read such literature surveys. So far as I can make out, not one in twenty scholars in my field is appreciably aware of the problem.

I believe there's a scandalous underestimation of the net impact of a number of factors on what the usual crude pro/con tally of significant and nonsignificant results probes. What my colleague Lykken calls the "ambient noise," or "crud factor," is of unknown average value, but it can hardly be supposed to be less than, say, in correlation terms, Pearson r = .25 in the soft areas of psychology. "Everything is correlated with everything," and .25 is probably not a bad average value. Randomly chosen individual differences variates do not tend to correlate zero. Of course in real life, the experimenter is usually correlating variates that belong, at least common-sensically, to some restricted domain. We don't usually do studies correlating social dominance with spool-packing ability or eye color. So a more realistic guesstimate of the crud factor, the expected correlation between a randomly chosen pair of variates belonging to a substantive domain, would be higher than that, maybe as high as .30.

Suppose an experimenter divides a group of subjects at the median on an input variable and includes all of these in the study as "high" and "low." If the input variable is normally distributed, then the difference in mean value between subjects in the high and the low group is around 1.6 sigma. If the crud factor is .30 in a broadly demarcated behavior domain, it would yield about a .5 sigma deviation expected difference on the dependent variable. If $N_1 = N_2 = 32$, we have a statistical power of around .50. So aside from the verisimilitude of the theory—it might in the extreme case have absolutely no truth in it—the research has about an even chance of getting a statistically significant result at $\alpha = .05$. Now of course the trend could be in either direction, so we might say there is about a .25 probability of getting a statistically significant result "in the right direction" that is, in the direction "predicted from theory," even though the theory has nothing to do with obtaining the effect. A random matching of theories with trend direction is, of course, set too low at $p = ½$, since the crud factor in most research areas is heavily concentrated in positive correlations and in some domains (e.g., ability, psychopathology) may yield a positive manifold.

When one surveys a body of research literature under these circumstances, a pro/con outcome tally of, say, 16:4 is far less favorable to the theory than it appears. There is an editorial bias in favor of significant results, partly due to Fisher's true but misleading dictum that H_0 cannot be proved, partly to power function problems, and perhaps to some feeling that they are more interesting or more informative. This editorial bias multiplies by an author's bias in submitting papers. Suppose the editorial bias favoring significant outcome (after excluding papers unacceptable for gross errors in design) were 2:1. and the author submission bias the same. This yields a pro/con bias of 4:1 in what studies finally appear in the literature. So an impressive box score of 16:4 (new theories' box scores rarely look better than that in *Bulletin* reviews) has arisen from a latent true pro/con outcome ratio about 1:1; that is, in reality about half of the experiments performed support the theory.

Now assume the investigators (having taken to heart the strictures of Cohen [1977]) have designed their experiments so as to achieve a power equal to .75, a bit lower than he recommends. Then the likelihood ratio L_t / L_0 of the theory against the crud factor is about an even 1:1. The true split pro/con (reported + unreported outcomes) being also even, the posterior probability of the theory on all evidence to date would be the same as the prior. Taking the prior on theories in soft psychology to be, say, $P \leq .10$ (their long-term survival rate is surely no better than that), the Bayesian posterior will then read $P(T/e) \leq .10$. This pessimistic but realistic computation is very different from the usual reviewer summary that a theory with 16:4 success/failure rate is doing rather well, is "quite promising," or "deserves further research"; in reality the posterior odds are 9 to 1 against the theory.

Now I don't mean to say that we all ought to be literally computing Bayes's theorem and numerical probabilities in this way (although some statisticians would say "Why not?"). Nor do I argue that at this point research on a 16:4 "hits" theory should be stopped. 1 only emphasize that the apparent 16:4 box score favoring the theory does not really favor it at all.

At this stage, another disturbing element appears. Suppose one accepts the philosophers' maxim "Do not make a mockery of honest ad hoc-ery," and the Lakatosian notion that we forbid the *modus tollens* arrow to be directed at the theory's "hard core" instead of at its protective belt. That's all very well, but the examples that Lakatos and others adduce are from astronomy, physics, and. chemistry, in which the hard core of the theory is clung to (despite a few prima facie falsifiers) because the theory has a lot "in the bank" already. We recognize the unwisdom of premature discarding because of anomalies that are apparent falsifiers. Such a policy does well in sciences where it's possible for a theory to get a lot in the bank early on, as with Kepler, Newton, Mendeleev, or Morgan. But in psychology, what is taken as having a lot in the bank is usually one of these 16:4 tallies.

The point is that understanding the logic and statistics of the situation—the asymmetry between corroboration (which is weak) and falsification (which is strong), the properties of the power function, and the fact that H_0 is always false if taken as a literal value—shows us that we get an illusion of having a lot in the bank empirically for theories that are extremely weak and that have as yet passed only feeble tests. So that "honest ad hoc-ery" is here being performed on a theory that we have very little reason to believe has appreciable verisimilitude on the basis of its early track record of 16:4 "successful" outcomes. This mess arises partly from the inherent difficulty of testing weak theories, but also from slavish adherence to significance testing as the research method.

For the reader who is wondering whether he follows the preceding reasoning, 1 can provide a short, simple litmus test. An objection occurs to what I have just said: "But surely there is some author and editor bias in submitting articles in the biological and physical sciences, and they're doing pretty well; so why are you making it out to be so terrible in psychology and sociology?" If that strikes you as a tough question, then you haven't fully got the point yet. The point is this: A selective bias in manuscript submission and editorial acceptance exerts its malignant effect via the widespread abuse of null hypothesis refutation, treated as if it were a powerful method of testing weak theories. Social scientists have a tendency to think that this bias in what subset of performed investigations reaches the published research literature is a sort of minor blemish on our research methods, perhaps suggesting mild social pressure (advice to editors) to change these habits and some slight modification of the way we think about the published corpus. That is a gross understatement of the case I am making here. The asymmetry between falsification and so-called confirmation in inductive logic is insufficiently appreciated by investigators and scholars. Combined with the relative feebleness of the hurdle that a theory has to pass if all we require is that "the girls be different from the boys," this asymmetry has the result that even a rather small bias in article submission, in addition to another small to moderate bias in acceptance, means that the tally of pro/con empirical outcomes on a particular theory should undergo some correction down toward the pro/con midline. Even if the bias were equally present in physics or astronomy, it would not have the catastrophic consequence it does for the social sciences because physicists and astronomers do not normally test theories by refuting the null hypothesis. In the rare instances when a significance test is used in physics, chemistry, or genetics, it is used in precisely the reverse of the way we use it in psychology and sociology, as I pointed out seventeen years ago (Meehl 1967). When the equivalent of a significance test is employed in physics, astronomy, chemistry, and most of genetics, it is employed to falsify the substantive theory by showing that the empirical results lie outside the range of instrumental and sampling error. Physicists, of course, were using the old "probable error" this way before R. A. Fisher was born. And it is worth noting that the invention of chi-square by Karl Pearson at the turn of the century was intended to be used in this way, namely, to measure "frequency discordance." That is, the question was whether the observed frequencies in a table of frequency distribution departed from that specified by the substantive theory. But when we have only a directional expectation, such as is generated by the weak theories of soft psychology and sociology, such a point prediction is not made. What we do is refute the null hypothesis and then take its contrary as being strongly corroborative of the theory, whereas in reality, it is only weakly corroborative. That means that significance tests are used in the opposite way in the physical sciences from the way they are used in the social sciences, except in those rare cases where the social sciences generate a sufficiently powerful model to make numerical point predictions or narrow interval predictions.

In physics it wouldn't matter much if a few investigators failed to send in their negative result papers. Given ten investigations of a theory prediction that such and such point values within such and such narrow tolerances should be found in the laboratory, and eight of those come out right, it would be an astounding coincidence if the theory had no verisimilitude. It is extremely difficult to explain eight of them coming out correctly, assuming low verisimilitude, whereas it is not nearly as difficult to explain that two of the ten depart significantly outside the tolerance. That is a totally different state of affairs from the case of refuting H_0 in a weak psychological theory, where the crud factor is available to explain all sorts of tendencies, and a box score of eight to two, even if wrongly taken at face value, does not speak strongly for the theory substance.

The physicist, chemist, or astronomer can put good money in the theory bank by a rather small run of successes because of the fact that they all involve point predictions or narrow interval predictions. And this money in the epistemic bank is what warrants physical scientists engaging in honest ad hoc-ery, lest a good theory with high verisimilitude be prematurely slain. The point is that a box score of 16:4 in psychology, given the bias in what appears in the journals, puts very little money in the bank, so that when combined with the low prior ratio on almost any theory in these fields, it means a posterior in Bayes' formula that is unimpressive.

I don't know the facts about selective reporting in the physical sciences, but it is obvious that a discipline in which a "negative result" means a significant deviation from a theoretically predicted point value constitutes very much stronger information than the failure to refute H_0 in the social sciences. For that reason, I would be surprised if the reluctance of authors to submit insignificant results, or the leaning of editors in favor of accepting significant results, were anywhere near as strong in physics as in psychology and sociology. Because of the problems of the statistical power function, even a falsification in psychology doesn't count as heavily as a measurement of a velocity falling outside the experimental error counts against a theory in physics.

One hesitates to paint such a bleak picture without having a clever and convincing "cure" up his sleeve, but, alas, I am unable to provide one. I do, however, have some constructive suggestions. An absolute precondition for improving matters with regard to the testing of theories and the early elimination of theoretical turkeys is "negative," to wit, to see that a bad problem exists. There is a widespread, massive intellectual inertia in my profession with respect to null hypothesis refutation as a tactic, witness the fact that the majority of psychologists in the soft areas continue to proceed unapologetically in this way, despite numerous articles (going back almost a quarter of a century in highly visible journals) that have raised the problem from a variety of standpoints and a whole book dealing with the significance test controversy at a high level of philosophical and statistical sophistication that appeared almost a decade and a half ago (Morrison & Henkel 1970).

The first thing we must do is to increase the general awareness of the younger generation of teachers and researchers that, given the nature of our subject matter and the ubiquitous crud factor, the corroboration of weak theories by a moderately successful run of refutations of H_0 is a feeble research strategy. The widespread adoption of that strategy accounts in part for the long history of failed psychological and sociological doctrines, each of which gave the illusion of great promise in its early phase. People just don't want to face the unpleasant fact that the base rate of long-term survival of theories in the social sciences is very small, so that the combination of this with the peculiarities of the significance test do not objectively yield the degree of corroboration of substantive theories that is generally supposed they do.

SUBSTANCE AND SIGNIFICANCE

For scholars to get the full point, it is not sufficient that they understand about the crud factor and about the statistical power function in relation to the crud factor. It is crucial also to understand the difference between a substantive theory and a statistical hypothesis that is indirectly related to it. It is salutary to reflect upon the reason why null hypothesis testing was fairly successful in the area of its first application (agronomy), namely, the very small (negligible except to a professional philosopher) logical distance or difference in meaning between the counter–null hypothesis and the substantive theory. I find that few psychologists and sociologists are clear about that, which I think reflects the manner in which undergraduate (and even some graduate) teaching of statistics is conducted. In such instruction, no professor should introduce students to the idea of doing a significance test without first distinguishing between substantive theory and statistical hypothesis, and then going on to point out that in agriculture, where Fisher made his great contributions, there is essentially no difference between the statement "The fertilized plots yielded more bushels of corn" and the statement "Fertilizing causes more corn to grow." When we refute H_0 statistically (directionality being taken for granted in agronomy), we corroborate its alternative, the counter–null hypothesis, which is the first statement. The confidence with which H_0 is refuted is in essence identical with the confidence we are entitled to have in our substantive causal statement.

In contrast, when refuting the null hypothesis as a means of corroborating a complex structural (compositional), functional, or developmental theory of neurosis, or perception or social dominance or whatever, this quasi-identity between the content of what we prove by refuting H_0 and what we want to prove substantively does not exist. This is partly because of the nature of the subject matter, since psychological theories usually involve hypothetical constructs while the agronomy theory is essentially a first-level observational inductive statement, and partly because of the ubiquitous and non-negligible crud factor, which could be understood by a Bayesian as a sizeable box of viable competitive theories that go in the denominator of Bayes's formula. There is a surreptitious tendency to mentally subtract the significance level from one, so that the complement $(1 - \alpha)$ gets vaguely "attached" to our confidence in the theory. Nobody explicitly does this. But the presence of those double and triple asterisks in a table of *t*-tests or *F*-tests (Meehl 1978) produces a misleading degree of subjective confidence by an unconscious assimilation of this complement value $(1 - \alpha)$ to the probability of the substantive theory. People commit, without being aware of it, the fallacy of thinking "If the theory *weren't* true, then there is only a probability of .05 of this big a difference arising," when of course we are not entitled to say anything even vaguely approaching this.

If there were adequate appreciation of the relative feebleness of null hypothesis refutation as a theory tester, as well as of its malignant combination with manuscript submission and editorial acceptance policies to give a biased box score in the published literature, what then might be done constructively? I trust my comments will not be misconstrued to mean that I disagree with the desirability of adequate statistical power as proposed forcefully by Cohen (1977), an important methodological thesis still not properly recognized by social scientists. It should be more widely emphasized that in order to set up a meaningful test of a substantive scientific theory, one needs, if employing significance testing at all, an adequate value of the power function. If more people met this requirement in conducting their research, then editors would not be presented with a dilemma of "keeping something out of the literature that should be known" while realizing that failure to refute the null hypothesis does not speak strongly against a particular substantive theory in the investigation because the power was too low. The APA Board of Publications should address itself to this question and adopt a strong policy.

I think it is scandalous that editors still accept manuscripts in which the author presents tables of significance tests without giving measures of overlap or such basic descriptive statistics as might enable the reader to do rough computations, from means and standard deviations presented, as to what the overlap is. In my view, it is inexcusable to present quantitative data in such a form that the reader is unable even to ask how many standard deviation units the experimental group was above the controls. Such data reporting is as incomplete as it would be not to mention which group intelligence test was used or how many degrees of freedom or where the sample was obtained. This is a gross defect in reporting scientific findings. Editors ought to arbitrarily reject such papers, so that members of the profession would come to take it for granted that measures of overlap and effect size must be presented. I don't wish to be dogmatic about what form should be used, although my own view is that both a metrical and a counting form should normally be employed and, if possible, the proportion of variance accounted for by the experimental factor.

It goes without saying that any statement of hits and misses by a cutting score should be accompanied by sufficient base-rate information and information about distribution shapes, so that the reader with clinical interest can make a meaningful assessment of how much has been achieved (Meehl & Rosen 1955). One of the best ways to reduce the illusion of scientific power that comes from writing "$p < .001$***" would be an accompanying table indicating—as would be very often the case with a test advocated for clinical purposes—that by using Fisbee's Projective Tennis Ball Test you can do 5 percent better at diagnosing schizophrenia in your clinic than you could by flipping pennies or guessing the base rate.

Ideally, of course, one would like to have stronger substantive theories, that is, theories that are capable of generating point predictions or relatively narrow interval predictions so that the significance test would have a meaning comparable to what it has in chemistry and physics or genetics. That is, we inquire whether our data depart significantly from the point value or the narrow interval that the theory demands. In such cases, the theory takes a high risk of falsification and consequently, if it succeeds in passing the hurdles, receives substantial corroboration. It should be realized that moderately strong theories can sometimes generate predicttions of patterns, of decreasing values of rank orders, of function shapes (e.g., that something will be "more or less ogival," even though not exactly the integral of Gaussian function), and so forth. I sometimes think that we social scientists suffer from a strange mixture of optimism and pessimism in this respect. On the one hand, we have been brainwashed by Fisherian statistics into thinking that refutation of H_0 is a powerful way of testing substantive theories. On the other hand, when urged to generate point or narrow-range predictions, we take it for granted that in the soft areas of psychology, it will be totally beyond our powers. Maybe the latter is the case, but I'm not convinced that it is. When one talks to applied mathematicians working in new fields like catastrophe theory, for instance, one hears that rather weak semiqualitative statements can sometimes be put together in ways that lead to rather specific quantitative predictions with a modicum of mathematical ingenuity. For example, one does not have to know the microstructure of a system in a way that points to predicting actual numerical values in order to be able to say that with increases in x, y must increase in a decelerated fashion up to the place that a third variable z equals 0, after which the derivative of y with respect to x increases.

There's an educational problem in psychology in this respect because we have a chicken and egg situation. Psychologists in the soft areas do not learn much of any mathematics (I don't count memorizing how to do an F-test as "mathematics"), and it is hard for the faculty to insist that they do so (especially if the faculty themselves don't know much math) because the student understandably wants to know what good it will do him since they don't use very much mathematics, but only t-tests, in his subdiscipline. But of course if nobody working in a given discipline knows any mathematics, they will never be able to find out whether it's possible to generate stronger semiquantitative predictions from relatively weak substantive theories.

It is interesting to ask whether research methods courses should explain the problem of higher-order interactions as a source of poor generalizability. There is nothing sinful about working with minitheories confined to a narrow domain. But a minitheory whose domain is narrowly restricted by, say, demographic variables is likely to be a rather poor minitheory. One suspects it would be possible, based on a careful literature survey, to write down a list of a dozen demographic factors and another dozen major methods factors that turn out to be the most nefarious in preventing strong generalizability of findings. It seems that there ought to be a research strategy that would take account of such expectable higher-order interactions, so that when a substantive theory of some process, say, social dominance or visual-perceptual learning or whatever, is proposed, there would be a recommended sequence for research studies aiming to test it, based upon our prior knowledge of the demographic and methods factors that seem to be most commonly a source of failures to replicate.

There is a problem here about the degree of densification in the nuisance parameter space that philosophers of science and statisticians should work on. It arises partly from the neglect of Brunswik's emphasis upon sampling situations as well as organisms. The "levels" of an experimental factor are not usually very problematic in agriculture, and reasonable levels are sometimes easy to select in fields like education. In domains where reasonable levels cannot be chosen on purely economic or ethical grounds, the problem of the distribution of patterns of experimental factors in a study of interactions becomes more difficult and complicated.

Finally, despite the absence of a rigorous definition of verisimilitude by the philosophers of science, I remain persuaded that some such concept is crucial in thinking about theory evaluation. Even a strong falsification (in which the auxiliary hypotheses are hardly in reasonable doubt) should be regularly viewed in fields like psychology and sociology as speaking strongly against the theory in its present form, rather than proving the theory to be deserving of instant execution and all further investigation of it abandoned forthwith. My emphasis upon falsification and the feebleness of H_0 refutation as a corroborator on the positive side does not mean that I disagree with the important qualifications and amendments of the original Popperian position by such critics as Lakatos.

I think some epistemological problems in social sciences cannot profitably be discussed unless the discussants are quite thoroughly familiar with concepts in philosophy of science. I note a tendency in some quarters to think that you can do philosophy of science quite casually. That is a grave mistake. My position is definitely not that all or most social scientists should know technical philosophy of science. What offends me is that from a state of philosophical ignorance, they advance methodological arguments that are inherently philosophical. My point is that if you are going to make use of what are in their very nature philosophical or epistemological arguments to defend or criticize a substantive or methodological scientific position (such as a certain research strategy or a preference for certain kinds of measuring instruments or a certain class of theories having properties in common), if you are going to employ philosophy of science for this purpose, you ought to know something about it.

If social scientists are going to proceed satisfactorily to some set of near-consensus conclusions on an accurate description of the state of affairs in a specified domain and what we should start doing instead, matters of philosophy of science and basic epistemology must either not come up because the nature of the topics being discussed does not inherently move to them or come up but be "settled" easily because the scientists already agree on at least an implicit philosophy of science, no matter how wrong-headed it may be. If the subject matter does force confrontation of philosophy of science and epistemology issues but the social scientists do not agree about those issues, they must possess technical competence—almost as much as the philosopher himself—before they can consider them fruitfully.

References

Carnap, R. (1936). Testability and meaning, part 1. *Philosophy of Science* 3: 420-471.

——. (1937). Testability and meaning, part 2. *Philosophy of Science* 4: 2-40.

——. (1946a). Remarks on induction and truth. *Philosophy and Phenomenological Research* 6: 590-602.

——. (1946b). Rejoinder to Mr. Kaufmann's reply. *Philosophy and Phenomenological Research* 6: 609-11.

——. (1948). Reply to Felix Kaufmann. *Philosophy and Phenomenological Research* 9: 300-304.

——. (1949). Truth and confirmation. In *Readings in philosophical analysis*, ed. H. Feigl and W. Sellers. New York: Appleton-Century-Crofts.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* 2d ed. New York: Academic Press.

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

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

Hull, C. L. (1943). *Principles of behavior*. New York: Appleton-Century.

McDougall, W. (1923). *Outline of psychology*. New York: Scribner's.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science* 34: 103-115.

——. (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.

——. (1983). Subjectivity in psychoanalytic inference: The nagging persistence of Wilhelm Fliess's Achensee question. In *Minnesota Studies in the Philosophy of Science, vol. 10. Testing Scientific Theories*, ed. J. Earman. Minneapolis: University of Minnesota Press.

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

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

Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.

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

——. (1958). *Semantics and necessary truth*. New Haven: Yale University Press.

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

Thurstone, L. L. (1935). *The vectors of mind*. Chicago: University of Chicago Press.

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

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

pdf by LJY April 2015