Meehl, P. E. (1989). Law and the fireside inductions (with Postscript): Some reflections of a clinical psychologist. *Behavioral Sciences and the Law*, 7, 521-550. [Reprinted in Meehl, *Selected philosophical and methodological papers* (pp. 440-480; C. A. Anderson and K. Gunderson, Eds.). Minneapolis: University of Minnesota Press, 1991.

Law and the Fireside Inductions (with Postscript): Some Reflections of a Clinical Psychologist

Paul E. Meehl, Ph.D.

Legislators and judges have relied upon the "fireside inductions" (commonsense, anecdotal, introspective, and culturally transmitted beliefs about human behavior) in making and enforcing law as a mode of social control. The behavior sciences conflict at times with the fireside inductions. While the sources of error in "common knowledge" about behavior are considerable, the behavior sciences are plagued with methodological problems that often render their generalized conclusions equally dubious. Legal applications of generalizations from experimental research on humans and animals in laboratory contexts often involve risky parametric and population extrapolations. Statistical analysis of file data suffers from inherent interpretative ambiguities as to causal inference from correlations. Quasi-experiments in the "real-life" setting may often be the methodologically optimal data-source. A postscript updates the original text and addresses seven additional topics: (1) abuse of significance tests, (2) failure to report overlap, (3) causal inference from correlation, (4) immediate transitions from group differences on psychological tests to "unfairness," (5) double standard of proof of generalizability, (6) social science in legal education, and (7) incompetent testimony by psychologists.

Paul E. Meehl, Ph.D., is Regents' Professor of Psychology, Professor of Psychiatry, and Professor of Philosophy of Science at the University of Minnesota. ~~*Correspondence and requests for reprints are to be addressed to Dr. Meehl, Department of Psychiatry, Box 392, Mayo Memorial Building,420 Delaware Street Minneapolis, Minnesota 55455.*~~

Editor's note: The complete version of this article was originally published in *The Journal of Social Issues*, Vol. 27, No 4, 1971, 65–100, and was part of the discussion, "Socialization, the Law, and Society," edited by June L. Tapp and published in Vol. 27, No. 2 (1971) of *The Journal of Social Issues*. The version reprinted here, with the kind permission of The Society for the Psychological Study of Social Issues, preserves most of the original material. Some minor changes have been made to conform the article to the style of this journal. Where sections have been omitted, editor's notes appear summarizing the contents. The postscript that appears at the end of the original text contains new material, and is summarized in a paragraph following the original abstract.

THE PSYCHOLOGY OF THE FIRESIDE

Lawmen will immediately see the point of my title, but for social science readers I should explain. The phrase "fireside equities" is legalese for what the legal layman feels intuitively or commonsensically would be a fair or just result (Llewellyn, 1960). Sometimes the law accords with the fireside equities, sometimes not; and lawyers use the phrase with derisive connotation. Analogously, by the language "fireside inductions" I mean those commonsense empirical generalizations about human behavior which we accept on the culture's authority plus introspection plus anecdotal evidence from ordinary life. Roughly, the phrase "fireside inductions" designates here what everybody (except perhaps the skeptical social scientist) believes about human conduct, about how it is to be described, explained, predicted, and controlled.

One source of conflict between the social scientist and practitioner of law—especially the legislator—is the former's distrust of common knowledge concerning human conduct and the latter's reliance upon this common knowledge. Such reliance is often associated among lawyers with doubts about the value of generalizations, arising from systematic behavioral science research involving quantification and experimental manipulation in artificial situations. Reliance upon "what everyone knows" (simply by virtue of being himself a human being) was hardly critically scrutinized prior to the development of the experimental and statistical methods of contemporary social science. This historical fact provides a built-in preference for the commonsense knowledge of human behavior embodied in positive law. But psychologists mistakenly suppose that the lawyers' continued reliance upon the psychology of the fireside is wholly attributable to inertia, and these misunderstandings warrant consideration. Without being honorific or pejorative, I shall use "fireside inductions" to refer broadly to those expectations and principles, largely inchoate although partially embodied in proverbs and maxims (e.g., "The burnt child dreads the fire," "Blood is thicker than water," "Every man has larceny in his heart," "Power always corrupts") arising from some mixture of (1) personal anecdotal observations, (2) armchair speculation, (3) introspection, and (4) education in the received tradition of Western culture prior to the development of technical social science method. It is not clear where nonquantitative, nonexperimental but psychologically sophisticated ideas, such as those of contemporary psychoanalytic theory and therapy, should be classified, but for the moment I will set this aside.

With my fellow psychologists I share a considerable skepticism concerning the fireside inductions. Even universally-held generalizations about the origins and control of human conduct should be subjected to (at least) *quantitative documentary* research and, where feasible, to systematic *experimental* testing. Obviously the degree of skepticism toward a dictum of commonsense psychology should increase as we move into those areas of social control where our efforts are hardly crowned with spectacular success. For example, there is no known system for the prevention or cure of crime and delinquency that is so strikingly successful that anyone can suggest we are doing so well at this social task that

it is hardly necessary to call our techniques into question, absent specific research that casts doubt upon them (Meehl, 1970c). That the psychological presuppositions underlying the criminal law should be subjected to merciless armchair scrutiny and quantitative research is not said *pro forma* but expresses a sincere conviction.

UNFAIR CONTROVERSIAL TACTICS SOMETIMES USED BY LAWYERS AND PSYCHOLOGISTS

Nor is this merely a platitude—"we need research"—that everyone accepts. One does come across rational, educated persons who disagree, at least when presented with concrete instances. I know, for instance, a very able law professor (formerly a practicing attorney) whose ignorance of the behavioral sciences was systematic and deliberate and who, although he regarded me highly as an individual intellect, made no secret that he thought most scientific research on law, such as quantitative studies on jury behavior, had little point. Over several months, I realized that he had a foolproof heads-I-win-tails-you-lose technique dealing with intellectual threats from the social sciences, to wit: If I introduced a quantitative-documentary or experimental study of some behavioral generalization having relevance to the law, the findings either accorded with his fireside inductions, or they did not. It they did, he typically responded, "Well, I suppose it's all right to spend the taxpayer's money researching that, although anybody could have told you so beforehand." If the results were *not* in harmony with the fireside mind-model, he refused to believe them! When I called this kind of dirty pool to his attention, he cheerfully admitted this truth about his debating tactics.

Without defending such illegitimate, systematic resistance to the inroads of behavior science data upon legal thinking, I direct my behavior sciences' brethren to some considerations that may render my law colleague's tactics less unreasonable than they seem at first glance. Some behavior scientists, particularly those ideologically tendentious and often completely uniformed with respect to the law, reveal a double standard of methodological morals that is the mirror image of my legal colleague's. They are extremely critical and skeptical about accepting, and applying in practical circumstances, fireside inductions but are willing to rely somewhat uncritically upon equally shaky generalizations purporting to be the rigorous deliverances of modern behavior science. A shrewd lawyer, even though he might not know enough philosophy, logic of science, experimental method, or technical statistics to recognize just *what* is wrong with a particular scientific refutation of the fireside inductions, may nevertheless be right in holding to what he learned at his grandmother's knee or through practical experience, rather than abandoning it because, say "Fisbee's definitive experiment on social conformity" allegedly shows the contrary.

Example: Punishment as a General Deterrent

Consider the threat of punishment as a deterrent, one of the most socially important and widely disputed issues relating behavior science to law. While I

have not kept systematic records of my anecdotal material (fireside induction!), the commonest reaction of psychologists upon hearing of my interest in studying law and teaching in the Law School, is a surprised, "Well, Meehl, I have always thought of you as a hard-headed, dustbowl-empiricist, quantitatively-oriented psychologist—how can you be interested in that medieval subject matter?" When pressed for an explanation of why they consider law medieval, my behavior science colleagues generally mention the outmoded and primitive (sometimes they say "moralistic") reliance of the criminal law upon punishment, which "is out of harmony with the knowledge of modern social (or medical) science." This kind of rapid-fire sinking of the lawyer's ship quite understandably tends to irritate the legal mind. However, the same psychologist who says punishment doesn't deter relies on its deterrent effect in posting a sign in the departmental library stating that if a student removes a journal without permission, his privilege to use the room will be suspended but his use fee not returned. This same psychologist suspends his children's TV privileges when they fight over which channel to watch; tells the truth on his income tax form (despite feeling that the government uses most of the money immorally and illegally) for fear of the legal consequences of lying; and drives his car well within the speed limit on a certain street, having been informed that the police have been conducting speed traps there. It will not do for this psychologist to say that as a citizen, parent, professor, taxpayer, automobile driver, etc., he must make such judgments upon inadequate evidence, but when contemplating the legal order he must rely only on scientific information. [Editor's note: Dr. Meehl goes on to provide examples of the faulty reasoning or evidence social scientists may offer to defend their dismissal of punishment as a deterrent, and notes that such practices may fuel the legal profession's skepticism about social scientists' pronouncements. Dr. Meehl then indicates that varying levels of sophistication in social science methodology may be needed to avoid misinterpreting data that relates punishment to crime rate.]

LEVELS OF SOPHISTICATION IN SOCIAL SCIENCE METHOD

This levels-of-sophistication question is of great importance in interdisciplinary work and in legal education. Any lawyer knows that having the more meritorious case does not guarantee winning it, a main interfering factor being skill of counsel. Differing levels of sophistication in any technical domain, even possession of a special vocabulary, often lead to misleading impressions as to who has the better of a theoretical or practical dispute. The parish priest can refute the theological objections of an unlettered Hausfrau parishioner. The priest, in turn, will lose a debate with the intellectual village atheist. C. S. Lewis will come out ahead of the village atheist. But when C. S. Lewis tangles with Bertrand Russell, it gets pretty difficult to award the prizes. This dialectical-upmanship phenomenon has been responsible for some of the friction between lawyers and social scientists, especially when the social scientist tendentiously presents what purport to be the findings of modern social science but is expressing the particular psychologist's, psychiatrist's, or sociologist's ethos or theoretical (ideological)

prejudices. Undergraduate sophistication sufficiently questions the efficacy of criminal law sanctions as a deterrent (although some college-educated legislators appear to be naive about this!), and recognizes the desirability of adequate statistics on comparable jurisdictions or within the same jurisdiction before and after a change in severity (or certainty) of penalty. However, to understand threshold effects, asymptotes, second-order interactions, nonlinear dependencies, rate changes at different points on a growth function—considerations hardly profound or esoteric—already takes us beyond the level of sophistication of many social and medical writers who have addressed themselves to legal problems. [Editor's note: Dr. Meehl then provides examples of overly quick, or unwarranted generalization on the part of social scientists. He observes that the presumed ineffectiveness of the death penalty as a deterrent does not justify "the rapid-fire dismissal of the general idea that increasing a penalty will be effective." He notes that such problematic generalizations reflect lack of knowledge or consideration of methodological and technical issues. In order to illustrate some of these methodological complexities, Dr. Meehl notes that increased punishment might decrease prohibited behavior under certain conditions, e.g., when apprehension is a near certainty and penalties are well publicized. Evidence is cited that supports this possibility, such as the decreased rate of military mutiny under the above cited conditions and variations in rate of crime when these conditions are and are not present. Dr. Meehl observes that in contrast, social scientists may attempt to countervail fireside inductions about punishment by invoking studies, such as those involving infrahuman animals, that are of doubtful application to everyday human affairs and require highly tentative extrapolations.]

THE CONCEPTS EMPIRICAL, EXPERIMENTAL, QUANTITATIVE

Rational discussion of the law's reliance on the fireside inductions may be rendered needlessly difficult by an unfortunate semantic habit as to the honorific word "empirical." Since I have myself been fighting a running battle with my psychological associates for some years against this bad semantic habit, I would dislike to see it accepted by legal scholars. The following methodological equation, often implicit and unquestioned, is being taken over by lawyers from behavior science:

Empirical = Experimental-and-Quantitative = Scientific

This equivalence is objectionable on several grounds. It is epistemologically inaccurate since there is a great deal of the empirical (i.e., arising in or supported by observations or experiences, including introspective experiences) that is neither experimental nor quantitative. Furthermore, the middle term assumes a false linkage because (1) not all experimental research is quantitative and (2) not all quantitative research is experimental. Third, several disciplines (to which hardly anybody would refuse the term scientific) exhibit varying amounts of experimental manipulation conjoined with varying amounts of the qualitative/quantitative

dimension, e.g., astronomy, ecology, botany, human genetics, paleontology, economics, meteorology, geography, historical geology, epidemiology, clinical medicine.

What is an experiment? I am not prepared to give an exactly demarcated definition of the term. Roughly, an experiment is a systematic, preplanned sequence of operations-cum-observations, the system of entities under study being relatively isolated from the influence of certain classes of causal factors; other causal factors being held quasi-constant by the experimenter; and still others manipulated by him, their values either being set for different individuals in the system or changed over time at the experimenter's will; and output is recorded at the time. Some (under Sir R. A. Fisher's influence) would add, but I would not, that remaining causally efficacious factors (known, guessed, or completely unidentified but assumed to exist) must be rendered noncorrelates of the manipulated variables by a randomizing procedure permitting their net influence to be estimated (statistical significance test).

This definition says nothing about apparatus, instruments, measurement, or even being in the laboratory. I disapprove of stretching the word "experiment" to include clinical or sociological research based upon ex post facto assessment processes, entering files of old data, naturalistic observations in the field or in public places, and so forth. But Campbell's "quasi-experiment" is useful to denote a subset of these possessing certain methodological features that render them relatively more interpretable (Campbell, 1969). The word "experiment" has become invidious because biological and social scientists tend to denigrate nonexperimental sources of knowledge (such as clinical experience, analysis of documents, file data, or the fireside inductions). Then, by equating "experimental" with "empirical" with "scientific," they often imply that any knowledge source other than experimental is methodologically worthless (armchair speculation, appeal to authority, metaphysics, folklore, and the like). But the fireside inductions *are empirical*. No logician would hesitate to say this. Their subject matter is the domain of empirical phenomena, and one who invokes a fireside induction will, when pressed to defend it, appeal to some kind of experience which he expects the critic will share with him, whether personally or vicariously.

Even the traditional law review article which traces, say, the development of a juridical concept like "state action" or "substantive due process" through a historical sequence of appellate court opinions is empirical, since its subject matter is the verbal behavior, recorded in documents, of a class of organisms, and the researcher studies the changes in this verbal behavior over time. The presence of analytical discourse in such a traditional law review paper does not render it nonempirical, but to argue this is beyond the scope of the present paper (see Feigl & Meehl, 1974; Meehl, 1970b; Skinner, 1969, Ch. 6).

There are important differences between the traditional law review article and the kind of article we expect to find in *Law and Society Review*. But we have some perfectly good words, more precise and less invidious, to characterize the difference. For a study of files or documents utilizing the statistical techniques of behavior science we can say simply "statistical," a straightforward word that

means pretty much the same thing to most people and which is not loaded emotionally. If structural statistics (such as factor analysis or multidimensional scaling, see Meehl, 1954, pp. 11–14) are employed, we have the word "psychometric." Distinguishing the quantitative or statistical from the experimental dimension is particularly important in discussing methodology of research on law because—as in clinical psychology and personology—one research method in these fields is the application of statistical and psychometric techniques of analysis to documents (e.g., diaries, interview transcripts, jury protocols, Supreme Court opinions). It would be misleading to say that one "performs an experiment" if one plots a curve showing the incidence of concurring opinions over time in the behavior output of an appellate court, but it is equally misleading to say that a traditional law review article which draws no graphs and fits no mathematical functions but traces through a set of opinions over time with reference to the incidence of split votes and dissents, presented in ordinary text, is not empirical. Research does not cease to be empirical, or even behavioral, when it analyzes behavior products instead of the ongoing behavior flux itself.

Since the control of variables influencing a dependent variable is a matter of degree, situations arise in which one is in doubt as to whether the word "experiment" is applicable. But this is merely the familiar problem of drawing an arbitrary cut whose location matters little. For research designs methodologically more powerful than studying a slice of cross-sectional file data because we have changes over time in relation to a societal manipulation (e.g., amendment of a penal statute), we have the expression "quasi-experiment" (Campbell, 1969).

MENTAL TESTS AND SOCIAL CLASS: THE LEVELS-OF-SOPHISTICATION POINT EXEMPLIFIED

The sophistication-level effect is beautifully illustrated by the vexatious problem of interpreting socioeconomic differences on mental tests. I suppose the minimum sophistication level necessary in order even to put the interesting questions is that of understanding why and how intelligence tests were built and validated, including basic concepts of correlation, content-domain sampling, reliability, validity, developmental growth curves, etc., that one learns about in an elementary psychology class. Exposure to basic psychometric theory and multiple strands of validation data (Campbell & Fiske, 1959; Cronbach & Meehl, 1955; Jackson, 1969; Loevinger, 1957) should eliminate some common antitest prejudices and excessive reliance upon anecdotal refutations. (*Example*: "I knew a kid with an IQ of 196 who became a bum.") I find it odd, by the way, that some lawyers will pronounce confidently to me about what intelligence tests do and do not measure, when the pronouncer could not so much as define IQ, factor loading, or reliability coefficient. I cannot conceive myself asserting to a lawyer that "the Hearsay Rule is silly" without having at least taken a course or read a treatise on the law of evidence, where the rationale of the rule and its exceptions was discussed. But I am uncomfortably aware that some psychologists permit

themselves strong—usually negative—views about the law without knowing any thing about it.

This bottom level of psychometric sophistication would suffice for an employer to consider using psychological tests in screening job applicants. At a higher level, one thinks about the social class bias in tests. If one stops there, relying upon the class-score correlations as definitive proof of test bias, one will perhaps avoid using tests, either because one may "miss" some good candidates (a poor reason, statistically) or from considerations of fairness, justice, equal opportunity (a good reason, provided the psychometric premise concerning bias is substantially correct). Moving one step higher in sophistication, we realize that the SES-IQ correlation is causally ambiguous, and that the limitations of statistical method for resolving this causal ambiguity are such that no analysis of file data can tell us what causes what. No one knows and, worse, no one knows how to find out to what extent the SES-IQ correlation is attributable to environmental impact and to what extent it is attributable to genetic influence. This causal ambiguity, while rather obvious (and clearly pointed out over 40 years ago in Burks & Kelley, 1928) is, as I read the record, somewhat above the sophistication level of many sociologists and psychologists, who talk, write, and design experiments on the (implicit) assumption that social class is entirely on the causal input side of the equation.

The terrible complexity of this problem cannot be discussed here, but I have treated it elsewhere (Meehl 1969, 1970a, 1971a). I can briefly concretize it by reference to the Coleman Report (Coleman, Campbell, & Hobson, 1966) on equality of educational opportunity. In the course of an interdisciplinary discussion at the University of Minnesota branch of the Law and Society Association one law professor argued: Since the Coleman Report showed that the psychological characteristics of a student's peer group were more closely correlated with his measured ability and achievement than either the school's physical plant or the characteristics of the teaching staff, these "empirical data of behavior science" would indicate that the way to achieve equal educational opportunity should be mandatory busing to provide disadvantaged pupils with the presumably better stimulation from abler peers. Whatever the legal merits of mandatory busing in relation to de facto segregation, the methodological point is important and requires a level of psychometric sophistication a notch above my law colleague's. It is possible that the higher statistical correlation between peer-group attributes and student's academic level is attributable mainly to geographic selective factors mediated by the family's social class, rather than causal influence of peer-group stimulation. Parental intelligence, personality, and temperament factors are transmitted to the child in part genetically (no informed and unbiased person today could dispute this, but many social scientists are both uninformed and prejudiced against behavior genetics) and partly through social learning. If physical-plant characteristics and teacher characteristics are correlated with the biological and social inputs of the child's family only via the (indirect) economic-neighborhood-location and political (tax-use) factors, they will have a lower statistical relationship with the child's cognitive level than is shown by indicators of the cognitive level

of other children attending the same neighborhood school. Roughly, peer-group attributes happen to be a better (indirect) measure of average family and neighborhood causal factors—genetic and environmental—than teacher or physical plant attributes. The differential correlation would reflect more a psychometric fact (about the factor loadings and reliabilities of certain measures) than a causal fact (about peer-group influence).

Whether one analyzes these data by inspecting a correlation table or by more complex statistical devices such as regression equations or analysis of covariance, neither the crude zero-order relationships, beta-weights, nor sums of squares tell us much about the direction of causal influence. We cannot infer whether social stimulation from other students is causally more efficacious than having better qualified teachers or a newer, better lighted, cleaner school building. The correlations with peer-group attributes cannot even tell us whether the impact upon a lower-IQ child of being in a classroom with more bright, dominant, articulate, and intellectually self-confident children does more harm than good.

In the opinion of Judge Skelley Wright in *Hobson v. Hansen* (1967) one cannot find a single sentence indicating recognition of this methodological problem. I do not suggest that awareness of it would have led to a different result. But a single sentence obiter would surely occur somewhere in his careful, scholarly opinion of 109 pages had the judge thought of it or counsel argued it in connection with the Coleman Report's significance. One may feel, as I do, that the problems of racial discrimination and educational disadvantagement are so grave that the society should lean over backwards—within limits set by principles of distributive justice to individuals—to change things, since we are confronted with a frightful combination of gross inequities and a major social emergency. Under such pressing circumstances the adoption and implementation of policy cannot await definitive solution of difficult scientific questions, especially when the kind of controlled experiment or even semicontrolled quasi-experiment (Campbell, 1969) capable of yielding clear answers cannot be performed and statistical techniques presently available are not adequate for the purpose of unscrambling causal influence. The Coleman Report shows that minority-group children receive substandard educational treatment, and I for one am willing to call that discriminatory, ipso facto. What concerns me here is the legal generalizability of a causal inference methodology. The kind of reliance upon social science data found in *Hobson v. Hansen*, lacking adequate clarification of the concept "unfair discrimination" in relation to correlational findings, might produce some untoward results in other contexts where the interpretative principles would be difficult to distinguish. And if judges should become cynical about the trustworthiness of what psychologists and sociologists assert, we might be faced with a judicial backlash against the social sciences. One can hardly blame Judge Wright for making a flat statement that intelligence tests do *not* measure innate intelligence or for repeating the old chestnut that intelligence is "whatever the test measures (*Hobson v. Hansen*, 1967, p. 478)." The Coleman Report states flatly, with no hint that disagreement exists among psychologists on the highly technical and obscure issues involved, "recent research does not support [the] view" that

intelligence tests measure more fundamental and stable mental abilities than achievement tests (Coleman et al., 1966, p. 292). "Ability tests are simply [*sic*] broader and more general measures of education" follows at the same locus, again without the faintest whiff of doubt or qualification to warn the legal reader that this is a complex and controversial issue. We read further, "The findings of this survey provide additional evidence that the 'ability' tests are at least as much affected by school differences as are the 'achievement' tests," the causal language "affected by" being again unqualified, with no mention made that some psychologists would interpret the parallelism of ability-score differences and achievement-score differences as suggesting that the achievement differences are not primarily due to school differences but to intelligence differences (an interpretation that would fit well with the report's other findings). It is well recognized among psychologists concerned with the ex post facto design's deficiencies that the report is dangerous reading for the nontechnically trained, because of its pervasive use of causal-sounding terms: influence, affect, depend upon, account for, independent effect (*sic*). This causal atmosphere cannot be counteracted by the brief methodological sections, which contain the usual caveats. One wonders whether the report's authors were really clear about what regression analysis can and cannot do (Guttman, 1941, pp. 286–292). I am not arguing the merits, except to show that an unresolved scientific controversy exists which we psychologists have no right to sweep under the rug when we talk or write for lawyers and judges. If we present a distorted picture even in a good cause, implying that certain technical matters are settled when in fact they are obscure and controversial, the powerful forces of the lawyers' adversary system will, sooner or later, ferret out the secret. Could we then complain if the findings of social science were treated with less respect than those of chemistry, geology, or medicine by less tractable, more wised-up judges? [Editor's note: Dr. Meehl then provides an additional example of the complexities involved in analyzing correlational data, and concludes: "The problem of interpreting correlations and the influence of 'nuisance variables' is not a hairsplitting academic exercise, it is a major methodological stomach ache, arising in many legal contexts where social science findings are relevant to fair treatment or equal protection."]

AWAY FROM THE FIRESIDE AND BACK AGAIN

The levels-of-sophistication problem has a time component reflecting the stage of scientific knowledge. We psychologists should be cautious when an alleged principle of modern behavioral science appears to conflict with the fireside inductions. There are some embarrassing instances of overconfident generalization and unjustified extrapolation which were subsequently corrected by movement back toward the fireside inductions. It would be worth knowing how often such back-to-the-fireside reversals have taken place, and whether there are features of subject matter or methodology that render the counter-fireside pronouncements of social science prone to reversal or modification. Sometimes psychologists seem to prefer negating the fireside inductions, especially those embedded in

the received scholarly tradition (e.g., Aristotle should be beaten up wherever possible). For example, the experimental psychologists' revival of the constructs "curiosity" and "exploratory drive" seems strange to a nonpsychologist who has observed children or pets—or who remembers the opening sentence of Aristotle's *Metaphysics*, "All men by nature desire to know."

Consider three examples relevant to law as a means of social control. Traditional reliance upon punishment (aversive control) in socialization, both in suppressing antisocial conduct and in education, is horrifying to the contemporary mind. One reads about the execution of a 14-year-old for larceny in the 1700s, or Luther's description of his schooling, in which corporal punishment was not even confined to infractions of discipline, but was the standard procedure of instruction. If a child didn't give the right answer, he would be rapped on the knuckles with a rod. Research on white rats and, to a lesser extent, on human subjects led to the generalization that, by and large, punishment is a mode of behavior control inferior to reward (positive reinforcement). Thirty years ago, I was taught that the useful role of punishment was to suppress undesirable responses sufficiently (in the short run) so that alternative competing behaviors could occur, and the latter could then be positively reinforced. This is still a fair statement of the practical situation.

Following the publication of Skinner's epoch-making *The Behavior of Organisms* (1938), his student Estes's doctoral dissertation (1944), and especially Skinner's *Science and Human Behavior* (1953) and his Utopian novel *Walden Two* (1948), aversive control fell into extreme disfavor. These writings combined with the gruesome stories told to us clinical psychologists by adult neurotics about their aversively controlled childhoods to produce a rejection of both the general deterrent and rehabilitative functions of the criminal law. But this could be an illicit extrapolation, conflating the rehabilitative and general deterrent functions of the criminal law under the generic rubric "punishment." Supporters of general deterrence need not assume the same psychological process operates on deterrable persons as that involved when punishment is unsuccessful in reforming convicts. Only punishment as a reformer even approximates the laboratory model of an aversive consequence following emission of the undesired response. Furthermore, the criminal sanction is rather more like withholding positive reinforcement (given elimination of flogging and similar practices from the penal system), both fines and imprisonment being deprivations. The distinction is becoming fuzzed up by experimental work with animals because manipulations such as "time out" (during which the instrumental act cannot be performed because the manipulandum is unavailable, or a stimulus signals that reinforcement will now be withheld) has aversive (punishing) properties. Having neither expertise nor space for the details, I refer the reader to Honig (1966), especially the chapter by Azrin and Holtz, which should be read asking, "To what extent do the current experimental findings refute, confirm, or modify the fireside inductions concerning punishment?" How would the psychologist classify a statutory provision that threatens to deprive the citizen of, say, money that the citizen had never learned to expect, e.g., the Agricultural Adjustment Act (see *U.S. v. Butler*, 1936, p. 81). where Mr. Justice

Stone's dissent hinges partly on the semantics of coercion, which he argues must involve "threat of loss, not hope of gain." Can the experimental psychologist speak to this issue? I doubt it.

A second example concerns imitation. The folklore is that both humans and infrahumans learn by imitation. (The criminal law, as lawyers and psychologists agree, is invoked to handle trouble cases, where the normal processes of socialization have not been applied or have failed to work. I trust my selecting the notion of general deterrence as a fireside induction will not be misconstrued as a belief in its major socializing role, which I daresay no psychologist would care to defend.) Our policy concerning TV and movie presentation of social models of aggression or forbidden sexual behavior is influenced by our beliefs about imitation. Despite the related Freudian emphasis upon identification as a mechanism of character formation, when I was a student the tendency in academic psychology was to minimize the concept of imitation to the point of skepticism as to whether there was any such process at all. I was taught that the classic experiments of E. L. Thorndike on the cat (*circa* 1900) had demonstrated that, for infrahuman organisms at least, there was no learning by imitation. This alleged laboratory refutation was presented as an example of how scientific research had overthrown part of the folklore. The failure of Thorndike's cats to learn one particular problem-box task, under his special conditions of drive and so forth, was overgeneralized to the broad statement, "Infrahuman animals cannot learn by imitation." The received doctrine of scientific psychology became so well entrenched that a well-designed experiment by Herbert and Harsh (1944) was largely ignored by the profession (see Barber, 1961). But beginning slightly earlier with Miller and Dollard's *Social Learning and Imitation* (1941), a book that cautiously reintroduced the concept and made important conceptual distinctions as to kinds of imitation, the subject came to be restudied, especially by developmental psychologists in relation to aggressive behavior (see, e.g., Bandura & Walters, 1963; Megargee & Hokanson, 1970). A recent article by John, Chesler, Bartlett, and Victor (1969) makes it probable that Thorndike's negative dictum at the turn of the century was just plain wrong, even for *Felis catus*. *Point*: A lawyer in 1930 might have lost a cocktail party debate with an animal psychologist, but the lawyer would have been closer to the truth.

A third area important in such legal contexts as presentence investigation is that of forecasting behavior probabilistically. The fireside inductions say that you should rely heavily upon the record of an individual's past conduct. As I have argued elsewhere (Meehl, 1970c), it may be that a naive judge will (over the long run) make better decisions than one who knows just enough psychology or psychiatry to rely on medical or social science experts making an intensive study of the offender. The efficiency of actuarial prediction is almost always at least equal to, and usually better than, prediction based upon (purported) clinical understanding of the individual subject's personality (see references in Footnote 4 in Livermore, Malmquist, & Meehl, 1968; and Footnote 8 in Meehl, 1970c). Second, behavior science research itself shows that, by and large, the best way to predict anybody's behavior is his behavior in the past (known among my

colleagues as Meehl's Malignant Maxim). Hence the naive judge's reliance on the fireside inductions may yield better results than the intermediate-level sophistication, which knows enough to ask a psychologist's or psychiatrist's opinion, but does not know enough to take what he says *cum grano salis*, especially when clinical opinion conflicts with extrapolation from the offender's record.

The subtle interaction between levels of sophistication and the developing state of scientific knowledge is nicely illustrated by the Supreme Court's attitude toward statutes postulating inherited tendencies to mental deficiency and criminalism. In *Buck v. Bell* (1927) the court upheld the constitutionality of an involuntary sterilization statute for mental defectives, in an opinion famous for Mr. Justice Holmes's "Three generations of imbeciles is enough." The opinion naturally does not display sophistication about the varieties of mental deficiency, such as the distinction between high-grade familial deficiency (usually nonpathological, being merely the low end of the normal polygenic distribution) and the Mendelizing or developmental anomaly varieties, characteristically yielding a lower IQ, relatively independent of social class, and presenting differing eugenic aspects since some of them have no discernible hereditary loading and others a clearcut one. Without entering into such technical issues in genetics, the court came to what I would regard as the right result (see Reed & Reed, 1965). In *Skinner v. Oklahoma* (1942) involuntary sterilization of a habitual criminal was disallowed, again a right result in my view. The fireside inductions that underlay Oklahoma's statute are perhaps as strong and widespread for criminal or "immoral" tendencies as for mental deficiency. But the scientific data on inherited dispositions are much stronger in the one case than in the other, and *that* much social science knowledge the Court did possess. Suppose that a more refined taxonomy of delinquents and criminals should enable us to discover that some persons disposed to antisocial behavior get that way in part on a genetic basis (see Footnote 10 in Meehl, 1970c), although in most delinquents the etiology is social. A modified form of the fireside inductions underlying Oklahoma's unconstitutional statute would then be defensible, and a properly redrafted statute combining habitual criminality as a legal category with psychogenetic categories or dimensions could be upheld on the same grounds as Virginia's sterilization statute. But the court's new task would demand far more technical sophistication, especially given the ideological components that would saturate social scientists' opinions in the briefs, than was required for handling *Buck v. Bell* and *Skinner v. Oklahoma*.

DIRECT APPLICATION OF EXPERIMENTAL RESULTS TO LEGAL PROBLEMS: SOME RATIONAL GROUNDS FOR CAUTION

As a clinical practitioner who was trained at a hard-nosed, quantitatively-oriented, behavioristic psychology department (Minnesota has been called the "hotbed of dustbowl empiricism" by some of its critics), I sense a deep analogy between the problem faced by judge or legislator in balancing the fireside inductions against purportedly scientific psychological or sociological findings, and the perennial problem of how far we clinicians are entitled to rely upon our clinical

experience, lacking (or apparently contradicting) experimental or quantitative research. For myself as clinician, I have not been able to resolve this dilemma in an intellectually responsible way, although I have been steadily conscious of it and engaged in theoretical and empirical research on it for over a quarter century. So I am hardly prepared to clean up the analogous dilemma for lawyers. However, dwelling on this analogy may enable me to offer some tentative suggestions. There is a similarity in the pragmatic contexts of law and clinical practice, in that something will be decided, with or without adequate evidence, good or bad, scientific or anecdotal. A judge cannot leave a case undecided—although a logician could point out that law, being an incomplete postulate set, renders some well-formed formulas undecidable.

Let us strip the concept "scientific experiment" to its essentials, as I have tried to do in a rough meaning stipulation *supra*. Forget the usual images of glass tubing and electronic equipment operated by bearded gents wearing white coats in a laboratory. What, for instance, is the purpose of gadgetry? Scientific apparatus performs one of two functions. Either it plays a role on the input side, contributing to the physical isolation of the system under study and to the control or manipulation of the variables, or it facilitates the recording of observations (output side). We conceive a situation space whose dimensions are all physical and social dimensions having behavioral relevance. In research on human subjects, this set of dimensions will include such minor variables as the material of the experimenter's desk, since in our society the social stimulus value of an oak desk differs from that of pine.

If we are studying the impact of a psychoanalytic interpretation and design an experiment to smuggle this real-life phenomenon into the laboratory, what happens? We move in the situation space front the ordinary-life context of psychotherapy to the experimental context. This movement is in the interest of locating the system studied more precisely, because in the ordinary-life, nonlaboratory situation, the values of certain variables known (or feared) to have an influence are neither assigned by the investigator nor measured by him (with the idea of their influence being removed statistically).

There is no mystery about this, no conflict between a scientific and non-scientific view of the subject matter. The problem presented is quite simple—it is the solution that presents complexities. In order either to eliminate certain causal variables, hold them constant, or manipulate them, or to measure them, we move to a different region of the situation space. By (1) eliminating, (2) fixing, (3) manipulating, or (4) measuring (and "correcting for") the input variables, we intend to test generalizations as to what influences what, generalizations we could not reach in the natural, field, nonexperimental situation. The price we pay is that these generalizations are only known to hold for the new region of the situation space; their application in the ordinary-life context is an extrapolation.

This untoward consequence of the experimental method does not flow from tendentious, polemic formulations polarizing a scientific against a non-scientific frame of reference (empirical versus armchair psychology). Such locutions are

misleading, as they locate our methodological stomach ache in the wrong place. The problem can be stated in general terms within the scientific frame of reference. To concretize it: Suppose I am interested in the behavior of tigers. If I rely uncritically on anecdotes told by missionaries, hunters, native guides, etc., my evidence will suffer both on the input and output sides, i.e., from indeterminacy of the input and inaccuracy of the output. If I have accurate output data (e.g., carefully screened independent, convergent testimony by skeptical, reliable, scientifically trained observers, use of telescopic camera recording, high-fidelity tapes of the tiger's vocalizations), I may be able by such means to take care of the recording-accuracy problem (although hardly the recording-completeness problem). But I will still be troubled by indeterminacy on the input side. I don' know all of the inputs to this tiger when I am photographing him at a distance. I do not know his inputs an hour earlier when he was invisible to me, and I have reason to believe that those previous inputs alter his momentary state and change his behavioral dispositions.

Suppose, to get rid of this uncertainty on the stimulus side, I capture my tiger and put him in a zoo or bring him to my animal laboratory. I eliminate the influence of variations in the chirping of a certain bird as part of the tiger's surround. There is a sense in which I now don't have to be concerned about birds chirping, the only birds that chirp in a proper psychological laboratory being those that the experimenter himself introduces. But there is another sense in which I should be worried about the influence of bird chirping. An average level and fluctuation of bird chirps is part of the normal ecology of tigers in the wild. If I want to extrapolate my laboratory findings to the behavior of wild tigers, this extrapolation is problematic. The background of bird chirps may have a quantitative impact upon the tiger's behavioral dispositions, and perhaps upon his second-order dispositions to acquire first-order dispositions (Broad, 1933; Meehl, 1972).

In research on human subjects, it is frequently found that the influence of variable x upon variable y is dependent upon values of variable z, called by statisticians an "interaction effect." Interaction effects regularly occur whenever the sensitivity of the experimental design suffices to detect them. It is not absurd to suppose that, in human social behavior, almost all interactions of all orders (for instance, the influence of variable v on the interaction effect of the variable z on the first-order influence of variable x on variable y) would be detected if our experiments were sufficiently sensitive. When we liquidate the influence of a variable, either by eliminating it through physical isolation or holding it fixed, we are in danger of wrongly generalizing from our experimental results to the natural, real-life setting.

Law-trained readers unfamiliar with social science statistical methods may have found the preceding rather abstract. Suppose I am a developmental psychologist interested in the deterrent effects of punishment and I argue that the sanctions of the criminal law are inefficacious, relying on "Fisbee's classic experiments on punishment in nursery-school children." In Table I, I list differences that might be relevant in extrapolating from the laboratory study to the legal context.

TABLE I Extrapolating from Fisbee to Real-Life

Experiment	Criminal Sanction Against Larceny
Four-year-olds	Adults
Mostly upper middle class	Mostly lower and lower middle class
Mostly biologically normal	Numerous genetic deviates in group
Time-span minutes or hours	Time-span months or years
Social context: Subject alone	Social context: Criminal peer-group inputs
Reward: Candy	Rewards: Money, prestige, women, autonomy, leisure, excitement
Punishment: Mild electric shock	Punishment: Deprivations of above rewards (Punishment more like nonreward or time-out than strictly aversive stimulus onset)
Punished response emitted; experimenter aims at reform	Punished response not emitted by most; law aims at general deterrence
Subject's perception of situation: Who knows?	Subject's perception of situation: Who knows?

One need hardly be obscurantist or antiscientific in one's sympathies in order to be nervous about extrapolation in the joint organism-situational space from the region of the left-hand column to that of the right as ground for repealing a statute penalizing larceny. In the foreseeable future, lawmakers will unavoidably rely upon a judicious mixture of experimental research, quasi-experiments, informal field observation, statistical analysis of file data, *and* the fireside inductions. The legislator, prosecutor, judge, and public administrator—like the clinical psychologist—cannot adopt a scientistic purist posture, "I will not decide or act until fully adequate standards of scientific proof are met by the evidence before me," The pragmatic context forces action. In these matters, not doing anything or not changing anything we now do is itself a powerful form of action. When the fireside inductions are almost all we have to go on, or when the fireside inductions appear to conflict with the practical consequences of extrapolated experimental research or psychological theory, it would be nice to have some sort of touchstone as to pragmatic validity, some quick and easy objective basis for deciding where to place our bets. Unfortunately there is none.

Are Some Classes of Fireside Inductions More Trustworthy than Others?

This problem is so important in a society that has become sophisticated and self-conscious as to its own modes of social control that one might reasonably argue in favor of support for second-level empirical research aimed at developing a taxonomy of fireside inductions, enabling us to sort them into categories having different average levels of accuracy. We do not even possess a corpus of the explicit fireside inductions upon which our law relies. To ferret these out from statutes, appellate court opinions, the Restatements, and so forth would be a monstrous and thankless task, although I suggest that a random sampling of the documents might be worth doing. One might inquire, apart from whether the fireside inductions are corroborated by social science research, how many legal

rules, principles, and practices accord even with the fireside inductions of contemporary men. Most law professors will readily agree, I find, that much of our law is predicated on notions about human conduct that hardly anybody would care to defend. *Example*: A lawyer will advise a testator to leave a dollar to one of his children whom he wishes to disinherit, because courts have held that since a parent naturally tends to bequeath property to his child, failure to do so creates a legal presumption that the testator omitted him by inadvertence. I doubt that this presumption accords with the fireside inductions. Laymen agree with me that if a father is sufficiently *compos mentis* to write a valid will at all, his failure to mention a son is not attributable to forgetfulness. *Example*: One class of exceptions to the Hearsay Rule is "declarations against interest," when the declarant is unavailable as a witness. The unavailability creates a special need for hearsay, and the fireside induction is that the declaration's having been made against interest renders it more trustworthy. But the interest is required to be pecuniary, not penal (Livermore, 1968, pp. 76–78: McCormick, 1954, pp. 546–551). Surely our fireside inductions do not suppose that a man is more likely to be careless or mendacious in admitting rape than that he borrowed money! Similar oddities have been noted in the Hearsay exception for "Admissions of a party-opponent," where a litigant's predecessor or joint tenant falls under the exception, but not a tenant in common or a co-legatee (McCormick, 1954 , pp. 523–525). I cannot imagine that the fireside inductions concerning motivations for accuracy would support these distinctions, which arose not from empirical considerations in the psychology of testimony but through formalistic intrusion of property-law metaphysics into the law of evidence.

A taxonomy of fireside inductions based upon their substantive, methodological, and psychological properties might permit a rough ordering of inductive types as to accuracy, comparing the fireside inductions in each category with social science generalizations available in the literature. If strong taxonomic trends existed among researched cases, we would have some basis for judging the probable trustworthiness of those unresearched. We could inquire whether the following methodological features of a fireside induction are associated with a higher probability of its being scientifically corroborated:

> Hardly anyone entertains serious doubts about the induction. Persons of different theoretical persuasions agree about the fireside induction almost as well as persons holding the same theoretical position. There is a consensus of the fireside that cuts across demographic variables such as education, occupation, social class, religious belief, ethnic background, and the like. Within the legal profession prosecutors, defense lawyers, law professors, and judges are in substantial agreement. Personality traits (e.g., dominance, social introversion, hostility, rigidity) are not appreciably correlated with adherence to the induction. The particular fireside induction involves an observation of actions, persons, or effects that occur with sufficient frequency so that most qualified and competent observers will have had an extensive experience as a generalization basis. The fireside induction deals

> with relatively objective physical or behavioral facts rather than with complicated causal inferences. The policy implications of the induction are such that nobody's political, ethnic, religious, moral, or economic ideology or class interest would be appreciably threatened or mobilized by its general acceptance in the society or by lawmakers' or administrators' reliance upon it in decision-making. Sophisticated armchair considerations do not reveal a built-in observational or sampling bias that would operate in the collection of anecdotal support or refutation of the induction. The induction is qualitative rather than one that claims to make quantitative comparisons despite the lack of a reliable measuring device.

I do not profess to know the relative importance of items in this list, and I can think of exceptions to any of them as a touchstone. Thus, for example, the height of a person's forehead is a relatively simple physical fact. But judgments of forehead heights of prison inmates by their guards were shown[1] to be correlated with guards' estimates of their intelligence, whereas objectively measured forehead heights were not. (*Explanation*: The erroneous fireside induction that a low forehead—low-brow—indicates stupidity led guards' perceptions and/or memories of this simple physical feature to be infected with their behavior-based estimates of a prisoner's intellect.)

And Are Some Experimental Findings Safer to Extrapolate?

One asks here about the comparability of two groups of organisms as to species, developmental period, and status. What motives, what rewards and punishments, what time relationships are shared between the experimental context and the natural setting? An indirect lead as to extrapolability can sometimes be picked up from the experimental literature itself, asking, "To what extent do the experimental findings replicate over a variety of species, drives, instrumental responses, rewards, and punishments?" If we cannot generalize within the laboratory, moving to the field is presumably risky. How sizable are the relationships? A social scientist who countervails a lawyer's fireside induction by extrapolating from psychological research yielding two correlation coefficients $r = .25$ and $r = .40$ is just plain silly; but an unsuspecting lawman, overly impressed with social science statistical methods, might be taken in. *Point*: Random sampling errors aside, this correlational difference represents an increment of less than 10% in variance accounted for, and could easily be liquidated by moving to a not very distant region of the situation space.

To have the best of both worlds, one would want accurate recording on the output side and proper statistical treatment, but with the situation being very similar to the one to which we wish to extrapolate. Accurate observations, accurate records (instead of memory and impressions), appropriate statistical

[1] I cannot trace the reference, which is from the late Professor Donald G. Paterson's lectures.

analysis are all attainable in the field or natural life setting, lacking experimenter manipulation of the input. I am therefore inclined to view Campbell's "Reforms as Experiments" (1969) not as a second-best substitute for laboratory investigation, but as often intrinsically preferable, because the situational-extrapolation problem is so grave that the scientific precision of laboratory experiments with college students or school children is largely illusory.

THE LAWMAKER'S DILEMMA

The legislator's, judge's, or administrator's situation is most comfortable when there is a sizable and consistent body of research, experimental and nonexperimental (file data and field observation data), yielding approximately the same results as the fireside inductions. While one may be scientifically skeptical even in this delightfully harmonious situation, in the pragmatic context of decision making, rule writing, policy adopting, etc., such rigorous skepticism can hardly lead to pragmatic vacillation. Some sort of action is required, and all we have goes in the same direction. The methodologically unsatisfactory situations can be divided into three groups, differing in degree rather than kind: (1) No quantitative or experimental evidence is available or readily collectable before action must be taken. Here we rely upon the fireside inductions, these being all we have. A healthy skepticism concerning the fireside inductions, engendered by the study of social science, makes us nonetheless uncomfortable. (2) We have a large-scale adequately conducted study in the field situation supplemented by file data from different jurisdictions varying in relevant parameters (e.g., offense rate, community socioeconomic indices); these field-observation and file-data results accord with theoretical concepts developed experimentally on humans and infra-humans; but the conclusion conflicts with the fireside. Such a massive and coherent body of information should countervail the fireside inductions, even those with the admirable properties listed above. It seems difficult to dispute this, since by including file data from the nonlaboratory setting to which we wish to extrapolate, we are in effect comparing two sets of anecdotal data, one of which has the methodological advantage of being based upon records instead of relying upon our fallible and possibly biased memories of observations gathered nonsystematically as regards representativeness of persons and situations. *Example*: If statistics show that accuseds released without bail pending trial have such a low incidence of pretrial criminal offenses or failure to appear for trial that the bail system has negligible social utility (combined with its obvious inequity to the poor), our fireside inductions to the contrary should not countervail. But a nagging doubt persists, since other relevant statistics (e.g., ratio of reported felonies to arrests) tend to support the fireside induction that some fraction of these defendants have committed further crimes during their pretrial freedom, and we cannot accurately estimate this fraction (see the excellent article by Tribe, 1970). A lively sense of the lawmaker's dilemma can be had by reading the Senate debate on the District of Columbia crime bill (*Congressional Record*, 1970). (3) The most difficult situation is that in which there is a collision between

a fireside induction having several of the good properties listed above, and a smattering of social science research that is strong enough to give us pause about the fireside inductions, but not strong enough to convince us. Thus, the research may not be entirely consistent from one investigator to another; or it comes only from the experimental laboratory and consequently involves considerable extrapolation in the situation-space; or, if a large-scale quantitative nonexperimental survey, it has the causal-ambiguity and variable-unscrambling difficulties intrinsic to such studies (Meehl, 1969, 1971a). One hardly knows what to suggest in such collision situations except the social scientist's usual "more research is needed."

CONCLUSION

Unavoidably, the law will continue to rely upon the fireside inductions. They should be viewed with that skepticism toward anecdotal evidence and the received belief system that training in the behavioral sciences fosters, but without intellectual arrogance or an animus against fireside inductions in favor of over-valued or overinterpreted scientific research. I can summarize my position in one not very helpful sentence since nothing stronger or more specific can be said shortly: In thinking about law as a mode of social control, adopt a healthy skepticism toward the fireside inductions, subjecting them to test by statistical methods applied to data collected in the field situation; but when a fireside induction is held nearly *semper*, *ubique*, *et ab omnibus* a similar skepticism should be maintained toward experimental research purporting, as generalized, to overthrow it.

POSTSCRIPT

I find, on reading what I wrote almost 20 years ago, that there is little to change on the basis of theoretical arguments or empirical evidence that have appeared since that time. There are, however, a few matters that fall under the general umbrella of social science research in relation to the "common sense" or "general knowledge" that lawyers, lawmakers, and judges must unavoidably rely on, that I did not discuss and that are at least as important as those that I did. Space limitations prohibit consideration of all the pros and cons; so my presentation may sound dogmatic. Perhaps the best way to view the text that follows is as raising questions, which, whether my answers to them turn out to be right or wrong, I think no informed person could deny are of legal importance.

Abuse of Significance Tests in Appraising Theories

It seems to be the fate of social scientists, in their impact on other professions, to begin by converting them to a procedure or substantive view that is slowly assimilated by the other group; and then the task of the social scientist is to emphasize the dangers of overdoing the lesson learned! I believe we have oversold statistical significance tests to law professors. Not that one should abandon them.

When a newspaper account of the short-term change in number of homicides fails to ask whether it is merely a "chance upward fluctuation," the psychologist must criticize it and hope that lawmakers and criminal justice system personnel do not overreact to something analogous to flipping pennies. My local newspaper publishes annually the achievement test scores of students in different schools, mislabeling the results as "grading the schools," without warning that some of the differences between schools, and between two testings in the same school, are probably chance variations. So it is appropriate to ask whether a difference between groups, or a trend line over time, is a genuine phenomenon, which we answer by doing a statistical significance test. For many questions in the social sciences, the first thing to ask about a set of data is whether their orderliness (trend, correlation, difference, change) is only apparent, i.e., could plausibly be due to random or chance factors. One sometimes says "tests of statistical significance are *necessary but not sufficient*," especially when the aim is to corroborate a causal theory for purposes of social action. In recent years there has been increasing criticism, both by statisticians and social science theorists, of the excessive reliance on statistical significance testing (proving that an observed sample difference could not plausibly be attributed to chance, so that the real difference, if we could observe the whole population, would be nonzero) as a way of proving that a substantive causal theory has verisimilitude (Bakan, 1966; Carver, 1978; Chow, 1988; Lykken, 1968; Meehl, 1967, 1978, [1990a], [1990b]; Morrison & Henkel, 1970; Rozeboom, 1960). Wherever possible, significance tests should be replaced by a statement of confidence intervals. That is, the numerical range within which one can have a stated assurance (*p* value) that the true population value lies. Preferable to that, when statistics are used to assess the verisimilitude of a substantive theory, efforts should be made to strengthen the theory sufficiently to permit prediction not merely of a nonzero difference but of the shape of a mathematical function, or the rank order of a set of groups, and—ideally—the numerical point value, as is done in the more advanced sciences such as physics, chemistry, and genetics (Meehl, 1978).

Failure to Report Overlap

When statistics are employed not primarily for evaluating a causal theory but for some technological purpose, such as justifying the use of a test in a military selection situation, or arguing for one remedial procedure over another in dealing with handicapped pupils, it is bad reporting to state the statistical significance level and not provide the reader with numbers indicating the overlap. I take a strong stand on this. I believe that it is unscholarly to submit such a paper, and that it is equally unscholarly that editors continue to accept such. There are legitimate disagreements as to the optimal way to express this matter of overlap, and which one is preferable depends on the pragmatic context. An obvious answer when there is doubt about how to express the overlap between two groups (say, one group treated with psychotherapy and the other with drugs, or one group of offenders paroled and the other incarcerated) is to report several of the generally

accepted overlap measures and let the reader take his pick. Whenever two groups are contrasted with respect to the impact of a procedure (therapeutic, educational, reforming), the investigator should state what percent of the one group reached or exceeded standard reference percentiles of the other group. Thus, we might read, "Seventy percent of those treated with Elavil exceeded the 50th percentile (median) of the control group." My own preference is to use standard reference points at the 10th, 25th, 50th, 75th, and 90th percentiles of the other group. Another measure increasingly accepted is the *Effect Size* advocated by Glass, McGaw, and Smith (1981) and Hunter (1982), in which the mean difference of the two groups is divided by the standard deviation of the control group, or sometimes by the composite of the two standard deviations. In case the two distributions are each close to normal, there is a well known statistic devised by Tilton (1937) called the *Tilton Overlap*. I repeat that I do not urge this as merely a kind of frosting on the scholarly cake, but as a *minimum necessity for adequate scientific reporting*. The sad fact that journal editors are sloppy about requiring it does not justify the common practice of merely reporting "the two groups differed significantly at the .05 level of probability." A law reader who knows some elementary statistics will find it instructive to draw a couple of normal curves which, with sample size 100, show a statistically significant difference at the .05 level. The reader will find that relying upon such a procedure of change, or such an instrument of selection, will do only a few percentage points better than one could do by flipping pennies. Many devices employed in clinical psychology are not cost effective for this reason, although in an industrial or military setting when N is large, and depending upon the *selection ratio* (= applicants/jobs), even a test of rather poor validity may pay off.

Causal Inference from Correlation

Doubtless every introductory course in social science, and every beginning statistics lecturer or text, informs the student that "correlation doesn't prove causality." This is a loose way of saying it, since if a correlation between two variables is not due to chance and can be replicated in subsequent samples, it does prove *some kind of causality* at work; what it does not prove, taken as it stands, is a direct causal connection between the two variables measured. Thus if IQ is negatively correlated with dental caries (which it is, or at least used to be in the old data), there must be something causing this relationship if it is statistically stable. But we do not know whether bad teeth lowers the IQ, as some dental hygienists argued in the 1920s from these correlations, or whether people with lower IQs don't take proper care of their teeth, or whether some third unmeasured factor affects both of these and they have no direct causal connection with each other. The last is the current interpretation, because when we partial out the influence of social class the correlation vanishes.

The ideal way to ascertain direct causal influence is to manipulate variables, which is why the experimental method is preferred over collecting statistics when it is applicable. Sometimes a social change, such as enactment of a criminal

statute, comes sufficiently close to an experiment to be illuminating as to causality (Campbell, 1969). Sometimes "experiments of nature" can play this role. We have several examples of police strikes, or absent or distant law enforcers, or the Nazi inactivation of the Danish police, which were followed immediately by a significant rise in crime. Lacking experiments and quasi-experiments, untangling the causal paths in a system of correlated variables is a complicated problem, about which there continues to be dispute among social scientists and statisticians. Lawyers should be familiar with the concept of *path analysis* which, while it is not usually capable of providing a solid gold affirmative argument for a certain causal understanding of a many variable system, is usually capable of *refuting* a particular causal interpretation (Duncan 1966; Werts & Linn, 1970; but see Li, 1975; Loehlin, 1987; Meehl, 1989; Shaffer, 1987). Sometimes it can at least provide plausibility considerations, as, for example, the IQ/tooth decay relationship described above. If that correlation goes away when you hold social class constant, it remains conceivable, but not plausible, to say that bad teeth lowers the IQ. If only three causal models are viable, and two of the three path diagrams can be clearly refuted, it is often reasonable for the policymaker to act in reliance on the sole unrefuted one, absent competitors.

Immediate Transition from Statistical Discrimination to "Unfairness"

One cannot conclude that a psychological test is "unfair" to a particular group (ethnic, geographic, religious, social class, age, sex) merely from the fact that the test shows significant group differences. This should be obvious, but for some reason the media, and some politicians and judges, seem unable to grasp it. If two groups do differ with respect to a social or psychological trait, then a valid test *should* show a difference between them. It is an empirical fact that many human traits, both of body and mind, exhibit small, medium, and even large differences between groups of individuals demographically specified. There are racial, national, social class, geographical, and sex differences in various abilities, interests, temperamental traits, susceptibility to diseases (physical and mental), and socially "neutral" traits such as the strange difference in the ability to taste the bitter synthetic chemical phenylthiocarbamide (PTC). The incidence of PTC "tasters" varies over the earth's surface from a low of 10% in some ethnic populations to a high of 80% in others, even though this trait has no biologically adaptive significance, as the substance involved does not exist in nature but was created by the chemist. One must distinguish the question whether a test as such is "unfair" (that it invalidly attributes differences between groups) from the question whether it is "valid and fair" for a trait that differs between groups because of a history of social discrimination. In the latter case the *test* is not "unfair," but social practices have been. The prevalence of medically diagnosable chronic alcoholism among Irish is nearly 20 times greater than among Jews. In fact, the only two variables that have been shown statistically predictive of alcoholism are (1) alcoholism in a first degree relative and (2) being of Irish extraction (Goodwin, 1981; Vaillant, 1983). Evidence from twin and adoption

studies proves that alcoholism involves a strong hereditary predisposition, contrary to what most of us were taught in undergraduate sociology classes. Psychometric or biochemical tests for alcoholism, or predisposition to it, would show a large "discrimination" between Irish and Jews, as would the statistics of D.W.I. convictions. Does this prove, or even tend to prove, that such tests are "biased against the Irish," or that the criminal justice system and highway department must have a pro-Jewish prejudice? Of course not. What to do about real differences arising from societal unfairness is a deep and complex problem at the interface between political and ethical theory; e.g., is it morally proper to perpetrate distributive injustice to present *individuals* as means of achieving a kind of "statistical" justice to *groups* composed of *other individuals*? This is obviously not a question on which psychologists possess any special expertise, but discussion of it is not helped by confusions about psychometric validity.

Double Standard of Proof of Generalizeability

In the original article I pointed out the extent to which lawyers, lawmakers, and judges rely upon the "fireside inductions," common knowledge available to people just because they have lived in the world, observed human behavior, and perhaps thought about their own behavior. I tried to emphasize evenhandedly that sometimes the fireside inductions are pretty good and should not be lightly discarded on the basis of an alleged scientific proof from the social scientist; but, on the other hand, it is important that lawyers be aware of the extent to which social science research that is not flimsy and tendentious does refute commonsense views our grandmothers cherished. The plain fact is that some fireside inductions are sound, others are unsound, and most are a mixture. I know of no way to find out which, when a dispute arises, except to collect facts in the systematic manner of the social scientist. But I sense that *sometimes* (I do not say usually) the legal profession imposes a double standard in this manner. I give only a single example that relates to the previous mentioned problem of psychometric validity. As I understand it, the courts have held that it is improper for a business concern to use an intelligence test for selection purposes absent clear proof that in that particular setting it has validity for the specific job involved. Now this sounds reasonable, but I submit that it isn't. The law relies on hundreds of "generalizations" about human conduct, about the generality of traits, about the trustworthiness of eye witnesses, about how ordinary people "reasonably" conduct their everyday affairs, that have not been subjected to any kind of objective validation. Many of these generalizations would, if critically studied, turn out to be either false, or at least not highly generalizable from one situation to another. For example, we routinely admit character testimony, presumed probative with respect to whether a party (or witness) would be likely to do so-and-so. But the research on trait generality (starting with the classic study by Hartshorne & May, 1928) shows such high behavioral specificity that some social psychologists can even assert (I think wrongly, but never mind) "there are no traits, there are only situations." By contrast, there are hundreds of research

studies, in a variety of settings, involving many thousands of civilians and military personnel, in a variety of kinds of jobs, which show that *proficiency at almost any kind of task will be correlated with the general intelligence factor*. If you don't like the "intelligence" label for the statistical factor, you can simply label it *g*, as is increasingly the practice among psychologists (Betz, 1986). Hawk (1986) estimates, on the basis of his review of these studies and a consideration of the effect sizes, that the saving in the U.S. economy achievable by using tests of *g* as a selective factor amounts to around $80 billion a year (but cf. Linn, 1986, who considers that inflated). It may be argued that the existence of ethnic and class differences in measures of *g*, when combined with the previous history of social unfairness, requires that some sort of commutative justice be done at the expense of distributive justice and economic inefficiency, hence measures of *g* are bad even when valid for the job. The point I wish to make is that the empirical grounds for believing in the cross situational validity of measures of *g* is far greater in amount, quality, and diversity than the grounds we have for believing probably 90% of the unresearched fireside inductions that a judge and jury rely on in the legal process. This is what I mean by saying that a kind of' double standard of evidence is being applied.

Social Science in Legal Education

I think it is both difficult and socially unnecessary to instruct law students in the details of social science research method, particularly the technical machinery of mathematical analysis whose interpretation is sometimes disputed among experts. Hardly any lawyers, even those who become professors, will be doing such research solo, without a co-worker who is a social scientist; and the important distinction between being a *critical research consumer* and a *new knowledge producer* should never be forgotten. I hold the same view with regard to the education of clinical psychologists who are vocationally oriented to becoming practitioners rather then university teachers and research investigators (Meehl, 1971b). It is not necessary to put law students through the algebra of Sewall Wright's equations for path analysis. Most of them wouldn't remember how to do it a couple of years later, unless they had continued to do research using it in the meantime. What is important is that they should know that there is such a technique as path analysis, and have acquired a strong readiness to ask the critical questions appropriate as a litigator, judge, or legislative committee member where a psychologist or sociologist is testifying about a matter of social causation. One can learn enough path analysis conceptually, without details of the mathematics, to think rationally about the matter—something that even a bright, reasonable person will have trouble doing if he or she is totally ignorant of the conceptual issues and metatheory involved. Grave mistakes in reasoning are made by U.S. senators who are not stupid or uneducated but who are sadly uninformed about matters of this particular kind. *Example*: Recently a senator arguing against federal financial aid to secondary schools (a subject on which I have no opinion) pointed to the fact that the average SAT score of high school seniors in Mississippi,

which has a low per capita support for education, exceeded that of high school seniors in Michigan, which has a high per capita support for education. In making this argument, he ignored the fact that a much smaller percentage of Mississippi students intend to go on to college, and they are the ones who take the test, which is not required of all high school seniors. All he was proving was a difference in the percentage of cases above a certain region on the normal distribution curve, self-selected for their educational plans. I am confident that a properly designed course in law school would sophisticate law students so that they would routinely think of asking this kind of critical question in the presence of such an argument, and that would be far more useful to them and to society than to teach them how to perform a Kolmogorov–Smirnov significance test or an analysis of covariance. "How to parse complex social causality" can be explained with numerical *illustrations* from the path-analysis literature, without trying to teach lawyers the computational procedures for *doing* one, or the underlying mathematical theorems.

Incompetent Testimony by Psychologists

Psychologists, like other professional experts, have benefited in prestige and income from the current litigation explosion. Only 20 years ago, after two days on the witness stand in a murder case, I was asked the "M'Naghten hypothetical," and the prosecutor objected on the grounds that I was not a physician. Today, so far as I know, every jurisdiction admits testimony by psychologists as to mental illness, competence to make a will, suitability as a parent in child custody cases, impairment of function due to brain injury, etc. As one who enjoys the courtroom scene and is pretty good at it, I can hardly find these social changes objectionable. Scientific integrity, however, compels me to comment on some unwholesome features of the testimony situation.

It is well known by members of my profession that psychology is a heterogeneous subject matter, probably more so than any other allegedly "scientific" discipline (Meehl, 1987). I have academic colleagues, both psychologists, who can hardly converse with each other about their work, because one studies the electrochemistry of the walleyed pike retina and the other writes about Jung's theory of dreams! It is not invidious or turf-protective to recognize a plain fact, that these qualitative differences in subject matter (and, correspondingly, research method) are associated with differences in scientific status. How firmly corroborated are the facts, generalizations, and explanatory theories of a psychologist's subdomain? How clearly and objectively defined are its leading concepts? How much consensus exists among "accredited" persons working in a domain? In my own specialty area (clinical psychology, psychometrics, behavior genetics) I know that the validity of diagnostic instruments, the factual support of theories, and the efficacy of therapeutic interventions varies from *high* and *clear* to *low* and *doubtful*. Unfortunately there are practitioners who either do not know these facts or choose to ignore them in practice, including forensic contexts. I have observed psychologists on the witness stand, and read trial transcripts and depositions that led me to

wonder how such a person could get through an accredited doctoral program without learning the rudiments of critical, scientific, quantitative thinking in which I was trained at Minnesota. I have concluded that there are numerous licensed practitioners who are, literally, *not competent* to evaluate data in a scientific fashion. Law readers can easily convince themselves of this by reading the excellent treatise by Ziskin and Faust (1988), "must" reading for any lawyer or judge who has to deal with expert testimony by psychologists. (If one lawyer has studied it and his opponent has not, the latter will probably be totally *crushed*, if the former puts on seasoned experts of the scientific kind.) A frightening eye-opener in the special area of child sexual abuse is Wakefield and Underwager (1988).

This sorry situation of "incompetent expertise," which I could explain sociologically if space permitted, presents a grave problem to the courts. Reliance on the opinions of experts, permitting questions and answers that our Anglo-American rules of evidence disallow for lay witnesses, presupposes that the expert *is* objectively "expert," that he or she knows more facts and thinks more incisively about them than a nonexpert could or would. The trial scenario is not a good forum for the resolution of complex technical issues involving scholarly disagreement. (Hence the usual emphasis on the expert's "qualifications," typically piled on to dazzle the jury far in excess of what is necessary to establish expertise.) A judge, in admitting expert testimony and instructing the trier of fact about it, naturally assumes that while experts may disagree, any expert knows the basic facts and tools of the trade, and knows how to reason about them properly. A plumber can be safely presumed to know what the most competent plumbers know about plumbing; ditto an orthopedic surgeon, accountant, or electrical engineer. *In the "soft" areas of psychology (clinical, counseling, community, social, personality, developmental) this cannot be safely presumed*.

I do not conclude from this that expert psychological testimony should be disallowed, although I admit that a case can be made for that conclusion. I do believe that trial judges should feel free to exclude some of it, when it would not be reversible error to do so. An important kind of expert testimony should consist of a scholarly showing that no trustworthy expertise exists (either side!) in certain areas. But because of widespread scientific incompetence among practitioners, such critical testimony will collide with the customary legal standard expressed by, "Doctor, is it generally held by your profession that…." The correct answer to this question is often, "*Yes, it is generally held, but erroneously*." If our evidentiary rules do not permit this critical consensus-challenging role for the expert, then the idea of greatly restricting areas of psychological expertise (e.g., to straight actuarial generalizations, analogous to an insurance actuary's testimony concerning the empirical numbers appearing in life tables in a wrongful death action) becomes, regrettably, more appealing. I find myself ambivalent on this score, partly because I cannot persuade myself that the insanity defense to a criminal charge should be liquidated, although my views as to its needed reform are extremely radical and arguably open to constitutional objections (Livermore & Meehl, 1967; Meehl, 1983; but cf. Lykken, 1982).

REFERENCES

Azrin, N.H. & Holtz, W.C. (1966). Punishment. In W.K. Honig (Ed.), *Operant behavior: Areas of research and application* (pp. 380–447). New York: Appleton-Century-Crofts.

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

Bandura, A., & Walters, R.H. (1963). *Social learning and personality development*. New York: Holt, Rinehart & Winston.

Barber, B. (1961). Resistance by scientists to scientific discovery. *Science*, *134*, 596–602.

Betz, N.E. (Ed.) (1986). The *g* factor in employment (Special issue), *Journal of Vocational Behavior*, *29* (3).

Broad, C.D. (1933). *Examination of McTaggart's philosophy*. Cambridge, Eng.: Cambridge University Press.

Buck v. Bell. *United States Reports*, 1927, *274*, 200–208.

Burks, B., & Kelley, T.L. (1928). Statistical hazards in nature-nurture investigation. In *Twenty-seventh yearbook of the National Society for The Study of Education, nature and nurture. Part I: Their influence upon intelligence*. Bloomington, IN: Public School Publishing.

Campbell, D.T. (1969). Reforms as experiments. *American Psychologist*, *24*, 409–429.

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

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378–399.

Chow, S.L. (1988). Significance test or effect size? *Psychological Bulletin*, *103*, 105–110.

Coleman, J.S., Campbell, E.Q., & Hobson, C. (1966). *Equality of educational opportunity*. Washington: U.S. Government Printing Office.

Congressional Record, 1970, *116* (121), July 17, 1970.

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

Duncan, O.D. (1966). Path analysis: Sociological examples. *American Journal of Sociology*, *72*, 1–16.

Estes, W.K. (1944). An experimental study of punishment. *Psychological Monographs*, *57* (Whole No. 263).

Feigl, H., & Meehl. P.E. (1974). The determinism-freedom and body-mind problems. In P.A Schilpp (Ed.), *The philosophy of Karl Popper* (pp. 520–559). LaSalle, IL: Open Court.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in social research*. Beverly Hills: Sage.

Goodwin. D.W. (1981). *Alcoholism*: *The facts*. New York: Oxford University Press.

Guttman, L. (1941). An outline of the statistical theory of prediction. In P. Horst (Ed.) *The prediction of personal adjustment, SSRC Bulletin, No. 48* (pp. 251–311). New York: Social Science Research Council.

Hartshorne, H. & May, M.A. (1928). *Studies in deceit*. New York: Macmillan.

Hawk, J. (1986). Real world implications of *g*. *Journal of Vocational Behavior*, *29*, 411–414.

Herbert, M.J., & Harsh, C.M. (1944). Observational learning by cats. *Journal of Comparative Psychology*, *37*, 81–95.

Hobson v. Hansen. *Federal Supplement*, 1967, *269*, 401–519.

Honig, W.K. (Ed.) (1966). *Operant behavior: Areas of research and application*. New York: Appleton-Century-Crofts.

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: Sage.

Jackson, D.N. (1969). Multimethod factor analysis in the evaluation of convergent and discriminant validity. *Psychological Bulletin*, *72*, 30–49.

John, E.R., Chesler, P., Bartlett, F., & Victor, I. (1969). Observational learning in cats. *Science*, *159*, 1489–1491.

Li, C.C. (1975). *Path analysis*: *A primer*. Pacific Grove, CA: Boxwood Press.

Linn, R.L. (1986). Comments on the *g* factor in employment testing. *Journal of Vocational Behavior*, *29*, 438–444.

Livermore, J.M. (1968). *Minnesota evidence: Minnesota practice manual 22*. Minneapolis: University of Minnesota General Extension Division.

Livermore, J.M., Malmquist, C.P., & Meehl, P.E. (1968). On the justifications for civil commitment. *University of Pennsylvania Law Review*. *117*, 75–96.

Livermore, J.M., & Meehl, P.E. (1967). The virtues of M'Naghten, *Minnesota Law Review*, *51*, 789–856.

Llewellyn, K.N. (1960). *The common law tradition*: *Deciding appeals*. Boston: Little, Brown.

Loehlin, J.C. (1987). *Latent variable models*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports* (Whole Monograph Supplement 9).

Lykken, D.T. (1968). Statistical significance in psychological research, *Psychological Bulletin*, *70*, 151–159.

Lykken, D.T. (1982). If a man be mad. *The Sciences* (Journal of the New York Academy of Sciences), *22*, 11–13.

McCormick, C.T. (1954). *Handbook of the law of evidence*. St. Paul: West Publishing Company.

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

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

Meehl, P.E. (1969). Letter in "Input." *Psychology Today*. *3*(6), 4.

Meehl, P.E. (1970a). Nuisance variables and the ex post facto design. In M. Radner and S. Winokur (Eds.), *Minnesota studies in philosophy of science* (Vol. 4, pp. 373–402). Minneapolis: University of Minnesota Press.

Meehl, P.E. (1970b). Psychological determinism and human rationality: A psychologist's reaction to Sir Karl Popper's "Of Clouds and Clocks." In M. Radner and S. Winokur (Eds.), *Minnesota studies in philosophy of science* (Vol. 4, pp. 310–372). Minneapolis: University of Minnesota Press.

Meehl, P.E. (1970c). Psychology and the criminal law. *University of Richmond Law Review*, *5*, 1–30.

Meehl, P.E. (1971a). High school yearbooks: A reply to Schwarz. *Journal of Abnormal Psychology*, *77*, 143–148.

Meehl, P.E. (1971b). A scientific, scholarly, nonresearch doctorate for clinical practitioners: Arguments pro and con. In R.R. Holt (Ed.), *New horizon for psychotherapy: Autonomy as a profession* (pp. 37–81). New York: International Universities Press.

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

Meehl, P.E. (1978). Theoretical risk, 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). The insanity defense. *Minnesota Psychologist*, *32* (Summer), 11–17.

Meehl, P.E. (1987). Theory and practice: Reflections of an academic clinician. In E.F. Bourg, R.J. Bent, J.E. Callan, N.F. Jones, J. McHolland, and G. Stricker (Eds.), *Standards and evaluation in the education and training of professional psychologists* (pp. 7–23). Norman, OK: Transcript Press.

Meehl, P.E. (1989). Path analysis: Metatheory matters. Unpublished manuscript. [Published in:
Meehl, P.E., & Waller, N.G. (2002). The path analysis controversy: A new statistical approach to strong appraisal of verisimilitude. *Psychological Methods, 7*, 283-300.
Waller, N.G., & Meehl, P.E. (2002). Risky tests, verisimilitude, and path analysis. *Psychological Methods, 7*, 323-337.]

Meehl, P.E. [1990a]. Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108-141, 173-180. [Reference updated]

Meehl. P.E. [1990b]. Why summaries of research on a psychological theory are often uninterpretable. *Psychological Reports*, *66*, 195-244. Also in R. E. Snow & D. Wiley (Eds.), *Improving Inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 13-59). Hillsdale, NJ: Lawrence Erlbaum Associates, 1991. [Reference updated]

Megargee, E.I., & Hokanson, J.E. (Eds.) (1970). *The dynamics of aggression*. New York: Harper & Row.
Miller, N.E., & Dollard, J. (1941). *Social learning and imitation*. New Haven, CT: Yale University Press.
Morrison, D.E., & Henkel, R. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.
Reed, E.W., & Reed. S.C. (1965). *Mental retardation*: *A family study*. Philadelphia: Saunders.
Rozeboom, W.W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.
Shaffer, J.P. (Ed.). (1987). *Journal of Educational Statistics* [Special issue]. *12*(2).
Skinner, B.F. (1938). *The behavior of organisms*. New York: Appleton-Century-Crofts.
Skinner, B.F. (1948). *Walden two*. New York: Macmillan.
Skinner, B.F. (1953). *Science and human behavior*. New York: Macmillan.
Skinner, B.F. (1969). *Contingencies of reinforcement A theoretical analysis*. New York: Appleton-Century-Crofts.
Skinner v. Oklahoma. *United States Reports*, 1942. *316*, 535–547.
Tilton, J.W. (1937). The measurement of overlapping. *Journal of Educational Psychology*, *28*, 656–662.
Tribe, L.H. (1970). An ounce of detention. Preventive justice in the world of John Mitchell. *Virginia Law Review, 56*, 371–407.
U.S. v. Butler. *United States Reports,* 1936, *297*, 1–88.
Vaillant, G.E. (1983). *The natural history of alcoholism*. Cambridge, MA: Harvard University Press.
Wakefield, H., & Underwager, R. (1988). *Accusations of child abuse*. Springfield, IL: Charles C. Thomas.
Werts, C.E., & Linn, R.R. (1970). Path analysis: Psychological examples. *Psychological Bulletin*, *74*, 193–212.
Ziskin, J. & Faust. D. (1988). *Coping with psychiatric and psychological testimony* (Vols. 1–3). Marina del Rey, CA: Law and Psychology Press.

pdf by ljy March 2003