#169

In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What If There Were No Significance Tests?* (pp. 393–425) Mahwah, NJ : Erlbaum, 1997.

Chapter 14

The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions

Paul E. Meehl
University of Minnesota

Significance tests have a role to play in social science research but their current widespread use in appraising theories is often harmful. The reason for this lies not in the mathematics but in social scientists' poor understanding of the logical relation between theory and fact, that is, a methodological or epistemological unclarity. Theories entail observations, not conversely. Although a theory's success in deriving a fact tends to corroborate it, this corroboration is weak unless the fact has a very low prior probability and there are few possible alternative theories. The fact of a nonzero difference or correlation, such as we infer by refuting the null hypothesis, does not have such a low probability because in social science everything correlates with almost everything else, theory aside. In the "strong" use of significance tests, the theory predicts a numerical point value, or narrow range, so the hypothesis test subjects the theory to a grave risk of being falsified if it is objectively incorrect. In general, setting up a confidence interval is preferable, being more informative and entailing null hypothesis refutation if a difference falls outside the interval. Significance tests are usually more defensible in technological contexts (e.g., evaluating an intervention) than for theory appraisal. It would be helpful to have a quantitative index of how closely a theory comes to correctly predicting a risky fact, and one such index is proposed. Unlike widespread current practice, statistics texts and lecturers should clarify and emphasize the large semantic (logical) gap between a substantive (causal, compositional) theory and a statistical hypothesis about a population parameter that is derivable from the theory but does not derive it.

Any rational evaluation of the significance test controversy must begin by clarifying the *aim* of inferential statistics. The mathematics is usually quite rigorous or can be made so if desired; hence, dispute between "schools of statistics," or between critics and defenders of the conventional significance test procedure, is unlikely to concern the formalism, but instead involves some question of concept formation, epistemology, or context. This chapter considers (without claiming to settle) when it is and is not appropriate to use null hypothesis tests, examines some complexities we face when doing research in the social and life sciences, and suggests a corroboration index that may be useful in appraising theories. I emphasize the methodological and epistemological questions as being more important than the purely "statistical" ones that are conventionally stressed.

TECHNOLOGICAL VERSUS THEORETICAL CONTEXTS

The first distinction to make is that between the use of significance tests (or proposed alternatives) in a *technological* versus a *theoretical* context. Making this initial distinction does not presuppose that the succeeding analysis will reveal an important difference. It merely follows the sociologists' methodological principle that one should initially disaggregate, leaving open the possibility of reaggregation if the subdivision turns out not to matter; whereas, if one begins by aggregation, one may be throwing away important information that is not recapturable. It is a defect of statistics texts and classroom teachers of statistics that the division between technological and theoretical tasks is almost never made explicitly and, if made, is not examined in detail.

Beginning with technology, we have two broad kinds of questions: those dealing with purely *predictive* tasks (e.g., can the Minnesota Multiphasic Personality Inventory [MMPI] predict suicide risk better than the Rorschach?) and *intervention* problems (e.g., does a certain psychotropic drug help depressed patients more than cognitive psychotherapy?). Discerning no relevant difference between these two pragmatic tasks, I consider only intervention. (The predictive task is always a component of the intervention-appraisal task, because the latter compares outcome probabilities associated with interventions, including "doing nothing.") Depending on the utilities and disutilities (e.g., probability of death from an illness), it may be that inferring *any* influence, however small, from a certain intervention is so important that we would want to know about it. For example, I recently came down with *subacute bacterial endocarditis*, an infective lesion of the heart valve leaflets, which has around a 15% mortality rate (100% prior to antibiotics). Because the infective organism in my case was *Streptococcus viridans*, the antibiotics used were gentamicin and ceftriaxone. Doubtless the choice of antibiotics was based on comparative research, and the superiority of these two drugs in treating this grave illness may have been slight, although statistically significant. But when I am the patient, I care about even small differences in therapeutic efficacy among the antibiotics that might be used. However, if there were powerful counterconsiderations such as side effects, amount of discomfort, danger, time required, or even cost, then the size of a therapeutic difference might matter to the patient, the insurance company, and the physician. Even an alteration in life and death probabilities, if minuscule but involving a hundred times more expensive procedure, might affect some persons' willingness to accept a slightly higher risk.

What is the role of conventional significance testing in this technological context? One easy and obvious answer (which may be the right one, for all I know) is that the proper statistical procedure is to set up a confidence interval. Its protagonists argue that, if this interval is known with a specified probability, the null hypothesis H_0: $\delta = (\mu_1 - \mu_2) = 0$ either falls inside the confidence interval or falls outside of it; thus any practical question answered by a significance test also is answered by establishing a confidence belt, and the latter is more quantitatively informative. Everyone agrees that the formalism is identical for the two procedures, it being just a question of moving the denominator $SE_{\bar{d}}$ of a *t* test to the other side of the inequality for a confidence interval. When the difference in utilities of the two kinds of errors is sufficiently large, it has been cogently argued (e.g., by Simon, 1945) that it may be rational simply to look for the existence of an observed difference as to direction, because the observed difference between two means is an unbiased and maximum likelihood estimator of the true difference. Whatever may be the values of the two kinds of errors in a significance test, the best bet is that δ has the same algebraic sign as the observed $\overline{\mathrm{d}}$. (I should mention here that I am not a subjectivist Bayesian, and my discussion throughout postulates the existence of objective probabilities, whether they are physical frequencies or "rational, inductive logic" epistemic supports. Quantifiable or not, some claims of strong evidentiary support are wrong, e.g., astrology. That vexed issue is beyond the scope of this chapter.)

We all learned from Fisher that you cannot prove H_0 but you can (almost!) refute it. Although this is correct as a mathematical truth, the unelaborated assertion can be misleading. In practical contexts, when we have sufficient power $(1-\beta)$ so that there is not too big an asymmetry in the values of error rates α and β, we do want to make the "quasi-null" inference, not that H_0 as a precise point value is literally true, but that something close to it is. If we were to scrupulously refrain from saying anything like that, why would we be doing a significance test in the pragmatic context? The clinician's conclusion from a failure to refute H_0 (given high power) is: "These drugs do not differ (at all, or appreciably, or enough to matter), so I can use either—selected randomly or by other considerations (e.g., cost)." This point holds for both technological and theoretical uses of the significance test.

If we consider all of the statistical inferences or pragmatic "decisions" that an investigator makes in the course of a research career, conclusions that a parameter θ lies within a confidence interval $[l_1, l_2]$ based on $(1-\alpha)$ will err in α

proportion of the occasions. Unless I am mistaken, this statement is correct regardless of where one stands on the controversy between Bayesians, Fisherians, Neyman–Pearsonites, and pure-likelihood people. I have heard statisticians explain patiently to psychologists that "δ has the value it has, so no 'probability' applies." Common as this remark is and plausible as it sounds, I think it is a mistake. Any application of a probability number involves, one way or another, what Reichenbach (1938) called a *posit*, employing that number to decide how we should think or act about the instant case. If there is *no* intelligible sense in which a *p* value "applies" to a particular instance, on the philosophical or metaphysical ground that the parameter we are interested in has objectively a certain definite value (although unknown to us), it is unclear why anybody would engage in significance testing or confidence belt estimation. Suppose I am asked to wager whether a thrown pair of dice will turn up a deuce. If the dice are not loaded and the shaking is thorough and fair, any statistician—of whatever "school"—would say it is reasonable to accept a bet on a deuce with odds of 1:6, and it would surely be advantageous to accept a wager if the house offered me, say, 7:1 odds. Assume I am a betting person, happy with those benevolent odds, and believe the house is honest. Would it matter whether I am to bet (a) before the dice are shaken, (b) while the shaking is going on, or (c) after the shaking is concluded but the cup is still inverted so we cannot see how the dice have in fact come up? I do not think any reasonable person would care to distinguish among these three cases as concerns personal betting.

Depending on whether you are a strict determinist at the macro level (cf. Earman, 1986), you could say that it was already determined how the dice would land before the shaking begins; but whether you are a determinist or not is irrelevant here. As long as I have not looked at the dice, what odds it is reasonable for me to take are unaltered by the meta-comment, "It doesn't make sense to talk about the *probability* of a deuce, because we have stopped shaking the dice and either there is a deuce under the cup or there isn't." Furthermore, the argument that if something is objectively the case (although unknown to us), therefore, it is meaningless to talk about probabilities would apply equally strongly whether we are dealing with a problem of estimation or a problem of significance.

I think the confusion here comes from failure to make Carnap's (1945) distinction between the two kinds of probability. Probability$_1$ is *epistemic*, the assertion of it takes place in the metalanguage, its subject matter is "inductive logic," and it is about someone's state of knowledge, about the relation of evidence to a probandum. Probability$_2$ is a relative frequency; the assertion of it occurs in the object language (e.g., of physics, genetics, psychometrics). There is usually no practicable algorithm for computing probability$_1$. That is why scientific judgments or judgments in courts of law or in ordinary life allow a certain leeway within the domain of the rational. Whether an algorithm for inductive inference in complex cases (such as the evidence that Jones committed the murder) can ever be formulated is in doubt, but most logicians and philosophers of science think not, and I tend to agree.

It is tempting to conflate the inference relation between statistics and parameters with the relation between accepted parameter values and the substantive theory; and because the former is numerified (e.g., a Bayesian posterior, a confidence belt, a significance level), one tends to think the latter is numerified also, or (somehow) *should* be. For example, evolutionary theorists cite evidence from the facts of geographical distribution (Darwin's original basis), embryology, comparative anatomy, paleontology, and so on. Some of these fact domains are analyzed statistically, and recently available DNA comparisons are intrinsically statistical. Each of these diverse databases is capable of yielding *numbers*, quantifying various probability$_2$ values concerning their respective concepts. But suppose one then asks, "What is the scientific probability of the neo-Darwinian theory based on this heterogeneous body of empirical evidence?" Nobody has the faintest idea how to calculate a *numerical value* of $p_1(T \mid p_{2i}, p_{2j}, p_{2k} \ldots)$, an epistemic probability, not a physical relative frequency. The only probability number some might write for $p_1(T \mid e)$ is a human knower's strength of belief in T given evidence e, a "subjective probability" expressing that individual's betting odds.

However, in dealing with empirical investigations that permit specification of a physical model (events, persons, fruit flies, attributes) to which the formalism applies, although usually somewhat idealized and hence numerically only approximate, these two probabilities—one epistemic and one physical, one about the state of the human knower and one about the state of the external world—may be the same or very close. If I say that the objective statistical frequency of throwing a deuce is approximately .028, and I accept the proposition that unbiased dice are being thoroughly shaken and fairly thrown, then that same probability$_2$ number can be reasonably used to refer to my state of knowledge of the external world and, further, to warrant my accepting the corresponding betting odds. So it is quite appropriate in the pragmatic context to talk about probabilities despite one's belief—sometimes near certainty—that there is such and such an objective state of affairs that is at the time of betting (or choosing a treatment, or whatever) not directly known to us.

Significance tests are appropriate in technological contexts, such as evaluating the efficacy of interventions, but setting up confidence intervals is preferable. In scientific communication, authors should always present confidence intervals, so that readers who prefer them will have the information they desire without depriving the significance testers of what they want to know. I consider it atrocious malcommunication for psychologists to publish tables that do not even give the value of t or F, let alone the means and standard deviations for their data, and who only state whether or not significance was achieved at a given level (hence the "tabular asterisks" in Meehl, 1978). It forces the reader who wants to know, for example, the approximate overlap or the mean separation in relation to the standard deviations to do computations that the author should have done and the editor should have required.

THEORY APPRAISAL

Theory "appraisal" is preferable to theory "testing" because the latter phrase is too specifically tied to Sir Karl Popper's philosophy of science. The arguments I make here do not depend on the reader's sharing my neo-Popperian sympathies. Before getting to the role of statistics, we have to be clear about the sheer logic of appraising a scientific theory. It is always more complicated than most college professors make it sound. Adopting the following convenient notation (Meehl, 1978, 1990a):

T: Main substantive theory of interest;
A_x: Auxiliary theories relied on in the experiment;
C_p: *Ceteris paribus* clause ("other things being equal");
A_i: Instrumental auxiliaries (devices relied on for control and observation);
C_n: Realized particulars (conditions were as the experimenter reported);
O_1, O_2: Observations or statistical summaries of observations;

then the logical structure of a test of a theory is the conceptual formula:

$$(T \cdot A_x \cdot C_p \cdot A_i \cdot C_n) \vdash (O_1 \supset O_2)$$

where dots " $\cdot$ " are conjunctions ("and"), turnstile " $\vdash$ " is deductive derivability (entailment, "follows that . . ."), and the horseshoe " $\supset$ " is the material conditional ("If . . . then . . .").

The first thing to be clear about is that, given this structure, a "successful outcome" of the experiment obviously cannot clinch the truth of the theory T, even if all of the other components are granted, because the argument is in the logical form: "If p, then q; q; therefore, p." In deductive logic, there are four possible inference patterns when we say "If p, then q," and these are shown in Table 14.1. Treated as if it were a deductive inference, the theory appraisal formula is in the third figure of the mixed hypothetical syllogism, which is formally invalid, the

TABLE 14.1

Deductive Inference Possibilities for the Hypothetical Argument: If *p*, then *q*

Figure	Form	Name	Conclusion
I	If *p* then *q* *p* $\therefore q$	Modus ponens (“Establishing mode”)	Valid
II	If *p* then *q* $\sim p$ $\therefore \sim q$	Denying the antecedent	Invalid
III	If *p* then *q* *q* $\therefore p$	Affirming the consequent	Invalid
IV	If *p* then *q* $\sim q$ $\therefore \sim p$	Modus tollens (“Destroying mode”)	Valid

formal fallacy called “affirming the consequent.” It affirms the consequent (*q*) and concludes that the antecedent (*p*) is true. Unfortunately, it is the form of all scientific inference aimed at supporting a theory by verifying its observational consequences. (This is what led to the famous witticism by eminent philosopher of science Morris R. Cohen, “All logic texts are divided into two parts. In the first half, on deductive logic, the fallacies are explained. In the second half, on inductive logic, they are committed.”)

Notice that this occurs at two levels in the appraisal formula: It holds for the term on the right-hand side of the formula ($O_1 \supset O_2$); and it is a problem when we consider the formula as a whole, where the conjunction on the left entails the argument on the right. The difference is that the material conditional ($O_1 \supset O_2$) is *truth-functional*; it depends solely on the pattern of truth and falsity of the propositions and has nothing to do with their semantic content; “If Nixon is honest, I’ll eat my hat,” is a common-discourse example of the material conditional. Deductive entailment, symbolized by the turnstile (“$\vdash$”), hinges on the meaning and structure of the theoretical system we are investigating; if there were no theory, there would be no *reason* for expecting the O_1, O_2 conditional on the right.

It is this irksome truth of formal logic that gives rise to what philosophers call Hume’s problem, or the problem of induction. Likewise, it was this “scandal” that led Sir Karl Popper to say that we can never *justify* scientific theories, we can only make efforts to *refute* them. The conceptual formula for appraising a theory says: From the *conjunction* of the theory of interest *and* auxiliary theories that we rely on *and* all other things being equal *and* auxiliaries with respect to the instruments we use *and* conditions having been correctly reported, it follows that *if* we do (or observe) O_1, *then* O_2 will be observed. If O_1 and O_2 are observed, the right-hand conditional ($O_1 \supset O_2$) is satisfied, so the experiment "came out right" (i.e., as the theory predicted). If O_1 is observed but O_2 is not, we have ($O_1 \cdot {\sim}O_2$) factually, falsifying the conditional ($O_1 \supset O_2$), and thereby falsifying the left-hand conjunction. This inference is in the fourth figure of the syllogism; it is a valid logical form, the medieval logicians' *modus tollens* ("destroying mode"), made a byword of metatheory by Popper's emphasis on it as the basic testing process of science (O'Hear, 1980; Popper, 1934/1959, 1962, 1983; Schilpp, 1974). When strong efforts to refute a theory continue to fail so that the theory is not killed, we should not say it is confirmed or supported, but simply that it has, so far, survived tests. For the survival of tests, Popper used the word *corroborated*. Whether his is a thoroughly adequate analysis is beyond the scope of this chapter, so I confine myself to saying that the Popperian analysis in a pure form has, as yet, not seemed persuasive to the majority of philosophers of science, and I feel confident in saying it has not been accepted by most working scientists, although of course neither of these two failures to convert shows that Popper is wrong. The important point is that this state of affairs is not a matter of one's preferred philosophy of science, but it is a matter of formal logic.

In addition to that formal fallacy, the structure of the appraisal formula has regrettable methodological consequences that are not matters of taste or philosophy of science (such as whether one is a realist or an instrumentalist or a fictionist, or whether one is a Popperian or an inductivist, or any of these interesting disputes), but also are a matter of formal logic. The negation of a conjunction is formally equivalent to the disjunction of the negations. Thus, having observed O_1, when we observe the falsification of O_2, then the falsified truth functional relation represented by the horseshoe in ($O_1 \supset O_2$) falsifies the conjunction on the left. But,

$${\sim}(T \cdot A_x \cdot C_p \cdot A_i \cdot C_n) \equiv {\sim}T \vee {\sim}A_x \vee {\sim}C_p \vee {\sim}A_i \vee {\sim}C_n$$

where " $\vee$ " means "or." Thus, although we intended to appraise only the main substantive theory T, what we have done is to falsify the *conjunction*; so all we can say "for sure" is that *either* the theory is false, *or* the auxiliary theory is false, *or* the instrumental auxiliary is false, *or* the *ceteris paribus* clause is false, *or* the particulars alleged by the experimenter are false. How confident we can be about the falsification of the theory of interest hinges upon our confidence in the truth of the other components of the conjunction.

To think clearly about this, one must recognize the distinction between the *substantive theory T* (causal, or compositional, or both) being appraised and some *statistical hypothesis H** that supposedly flows from it. Hardly any statistics textbooks and, so far as I have been able to find out, hardly any statistics or psychology professors lecturing on this process bother to make that distinction, let alone emphasize it. A substantive causal theory, such as Festinger's theory of cognitive dissonance or Meehl's theory of schizotaxia, consists of a set of statements about theoretical entities that are causally connected in certain ways, and these statements are *never* equivalent to a mere assertion about such and such a population parameter. Despite this obvious conceptual difference between T and H^*, there is an almost irresistible temptation to move from a small p value in a significance test, via a high confidence that H^*: $\delta > 0$, to a (similarly high) confidence that the substantive theory T (which entailed that nonzero directional δ) is true, or at least has high verisimilitude. The trouble is that the directional nonzero δ can be derived from other theories than T, and in the "soft" areas of psychology a sizable number of those competing theories have plausibility.

An Example of Theory Appraisal. How would the theory appraisal formula apply in a research situation? To illustrate briefly, suppose we wish to test Meehl's theory that schizophrenia is the decompensated form of schizotaxia, a neurological disorder due to an autosomal dominant gene (see, e.g., Meehl, 1990d). Examining clinically normal parent pairs of carefully diagnosed schizophrenic probands, we predict that the parent showing "soft" neurological signs (e.g., SPEM eye-tracking anomaly, adiadochokinesia) will have a higher MMPI schizotypy score than the neurologically normal parent. The logical structure, much abbreviated here, reads:

T (theory of interest): Meehl's schizotaxia dominant gene theory.

A_x (auxiliary theory): For example, most schizotypes are not decompensated, they appear "normal."

C_p (*ceteris paribus*): No other factor (e.g., test anxiety in parents worried about their own mental health) exerts an appreciable influence to obfuscate the main effect.

A_i (instrumental auxiliary): For example, the MMPI is valid for psychological schizotypy; the soft signs are valid for schizotaxia.

C_n (realized particulars): For example, the probands were correctly diagnosed in accordance with *Diagnostic and Statistical Manual* (*DSM*) criteria, as alleged.

O_1 (first observation): Subset S_1 of parents have soft neurology pattern and spouses do not.

O_2 (second observation): Subset S_1 of parents each have MMPI score > spouse.

If we had no theory (Meehl's or some other, competing one) there would be *no reason* for expecting that the higher MMPI-scoring spouse would have difficulty tracking a moving visual target (the SPEM eye-tracking anomaly). If the conditional ($O_1 \supset O_2$) is empirically verified as predicted, we have corroborated *T*. If not, if we find factually that ($O_1 \cdot \sim O_2$), *T* has been discorroborated. How strongly, either way, depends on how much confidence we have in the left-hand conjuncts other than *T*. Corroboration and discorroboration are matters of degree.

THE CRUD FACTOR

The causal and compositional structure of mind and society (including biological matters like genetics) being what they are, almost all of the variables that we measure are correlated to some extent. In the behavioral sciences, the saying "everything correlates with everything," gives rise to what David Lykken has termed the *crud factor* (Meehl, 1990e). Some readers have taken Lykken and me to be referring to the crud factor as a source of random sampling error, and therefore somehow connected with the values α , β or their ratio, hence correctable by refining the statistics, converting to Bayesianism, or some other statistical ploy. This is a complete misunderstanding of our position. *The term "crud factor" does not refer to statistical error, whether of the first or the second kind*. The crud factor consists of the objectively real causal connections and resulting statistical correlations that we would know with numerical precision if we always had large enough samples (e.g., a billion cases) or if we had measured all of the members of the specified population so that no sampling errors (but only errors of measurement) remained. In purely correlational studies, the crud factor is ubiquitous. In experimental studies, where we randomly impose a manipulation, the true $\delta = \mu_1 - \mu_2$ or ρ may be zero. If so, the class of explanatory theories usually will be smaller; but the non-null relation is still not semantically equivalent to the substantive theory that entails it. (Lykken and I disagree about the pervasity of crud factor in controlled experiments.)

The ubiquitous crud factor in social science is what *empirically* corresponds to the *formal* invalidity of the third figure of the implicative syllogism (see Table 14.1). The whole class of theories that could explain an experimental finding is what makes it impossible to deduce *T* from the "successful" observational out come. This is a logician's or an epistemologist's problem; it is not a statistician's problem in the sense of something answerable by proving theorems in the formalism. Thinking of the class of substantive theories T, T', T'', . . . that could occur on the left of our appraisal formula as, say, a set of subject populations from which our sample could have been drawn, or a set of frequency distributions having various parameters μ, μ', μ'', . . . , σ, σ', σ'', . . ., is a totally wrong-headed notion. These hypothetical populations are what the *statistical hypotheses H, H', H''*, . . . are about. These Hs are numerical consequences of the substantive Ts, not their semantic equivalents. This is most clearly seen if we suppose the sign of δ to be known for sure, that is, as if $\alpha = \beta = 0$. In that epistemically delightful circumstance*, what does the investigator know* about the truth of T? The investigator clearly does not know for sure that T is correct, and this would be the case even if all of the other conjuncts on the left hand of the theory appraisal formula were known for certain. Critiques of null hypothesis testing (e.g., Bakan, 1966; Carver, 1978; Cohen, 1994; Hogben, 1968; Lykken, 1968; Meehl, 1967, 1978, 1990a; Morrison & Henkel, 1970; Rozeboom, 1960; Schmidt, 1996) are not primarily aimed at the two kinds of errors discussed in statistics books, and their criticisms cannot be rebutted by manipulation of the formalism.

I said that the crud factor principle is the concrete empirical form, realized in the sciences, of the logician's formal point that the third figure of the implicative (mixed hypothetical) syllogism is invalid, the error in purported deductive reasoning termed affirming the consequent. Speaking methodologically, in the language of working scientists, what it comes to is that there are quite a few alternative theories T', T'', T''', . . . (in addition to the theory of interest T) that are each capable of deriving as a consequence the statistical counternull hypothesis H^*: $\delta = (\mu_1 - \mu_2) > 0$, or, if we are correlating quantitative variables, that $\rho > 0$. We might imagine (Meehl, 1990e) a big pot of variables and another (not so big but still sizable) pot of substantive causal theories in a specified research domain (e.g., schizophrenia, social perception, maze learning in the rat). We fantasize an experimenter choosing elements from these two pots randomly in picking something to study to get a publication. (We might impose a restriction that the variables have some conceivable relation to the domain being investigated, but such a constraint should be interpreted very broadly. We cannot, e.g., take it for granted that eye color will be unrelated to liking introspective psychological novels, because there is evidence that Swedes tend to be more introverted than Irish or Italians.) Our experimenter picks a pair of variables randomly out of the first pot, and a substantive causal theory randomly out of the second pot, and then randomly assigns an algebraic sign to the variables' relation, saying, "H^*: $\rho > 0$, if theory T is true." In this crazy example there is no semantic-logical-mathematical relation deriving H^* from T, but we pretend there is. Because H_0 is quasi-always false, the counternull hypothesis $\sim H_0$ is quasi-always true. Assume perfect statistical power, so that when H_0 is false we shall be sure of refuting it. Given the arbitrary assignment of direction, the directional counternull H^* will be proved half the time; that is, our experiment will "come out right" (i.e., as pseudo-predicted from theory T) half the time. This means we will be getting what purports to be a "confirmation" of T 10 times as often as the significance level $\alpha = .05$ would suggest. This does *not* mean there is anything wrong with the significance test mathematics; it merely means that the odds of getting a confirmatory result (absent our theory) cannot be equated with the odds given by the t table, because those odds are based on the assumption of a true zero difference. There is nothing mathematically complicated about this, and it is a mistake to focus one's attention on the mathematics of t, F, chi-square, or whatever statistic is being employed. The population from which we are drawing is specified by variables chosen from the first pot, and one can think of that population as an element of a superpopulation of variable pairs that is gigantic in size but finite, just as the population, however large, of theories defined as those that human beings will be able to construct before the sun burns out is finite. The methodological point is that T has not passed a severe test (speaking Popperian), the "successful" experimental outcome does not constitute what philosopher Wesley Salmon called a "strange coincidence" (Meehl, 1990a, 1990b; Nye, 1972; Salmon, 1984), because with high power T has almost an even chance of doing that, *absent any logical connection whatever between the variables and the theory*.

Some have objected that the crud factor is not ubiquitous, or that it is usually negligible in size, so that H_0 is "almost true" in most psychological research. If the latter line of reasoning is adopted, one is now asking not whether $\delta = 0$, but what is δ's numerical nonzero value; and one should either restate the nonzero *numerical value* of the counternull H^* that "should rationally count" as a corroborator of the theory or use a confidence interval. I have been surprised to find psychologists seriously maintaining that H_0 is likely to be literally true in correlational studies in the "soft" areas of psychology (clinical, counseling, developmental, personality, and social psychology). It seems obvious on theoretical as well as common-sense grounds that the point value $\delta = 0$ can hardly ever be literally correct. Consider, for example, an experiment in which we draw the dichotomous variable *gender* and the continuous variables *speed* and *accuracy* of color naming from one pot and any arbitrary substantive personality theory T from the other pot. (If the reader objects to this absurdity, well and good, because that preposterous scenario is the one most unfavorable to the case Lykken and I are making.) Omniscient Jones knows the regression equation (in this instance also the causal equation, though they are not usually identical!) for males' color-naming score y to be

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$$

so that the mean color-naming score of males is

$$\mu_{\mathrm{m}} = \beta_1 \mu_{\mathrm{m}_1} + \beta_2 \mu_{\mathrm{m}_2} + \beta_3 \mu_{\mathrm{m}_3} + \ldots + \beta_n \mu_{\mathrm{m}_n}$$

The corresponding equation for females' means will be

$$\mu_{\mathrm{f}} = \gamma_1 \mu_{\mathrm{f}_1} + \gamma_2 \mu_{\mathrm{f}_2} + \gamma_3 \mu_{\mathrm{f}_3} + \ldots + \gamma_n \mu_{\mathrm{f}_n}$$

In order for the parameter $\mu_{\mathrm{m}} = \mu_{\mathrm{f}}$, so that H_0: $\delta = 0$ is a literally correct point value, either the β and γ weights and the means μ_{ij} of all of the determining variables (color vision, verbal fluency, aesthetic interest, number of conversations with mother about clothing, religiosity interest in seasonal changes in liturgical vestments, etc.—a long list) must be identical for boys and girls or, if not, differences in the μs must be delicately balanced with appropriate differences in the βs and γs so as to yield the same output means. I cannot imagine any social or biological scientist thinking that this could happen in the real world.

It would probably be worthwhile to know the average size of the crud factor, whether expressed as a correlation coefficient or a standardized mean difference, in various research domains in psychology; but researching this would be a rather thankless task. One easier study that should be done would be a large-scale archival sample in a couple of "soft" fields such as clinical and social psychology, in which the proportion of studies reaching a preassigned significance level α is plotted as a function of sample size, and a mathematically appropriate function fitted to those points. One could then answer the question, "What is the asymptote?" I record my confident prediction that the asymptote of this "significant result" function will be extremely close to $p = 1.00$. We do have some empirical evidence on this matter in a study by Lykken and myself (Meehl, 1990e); all of 105 pair-wise relationships between a hodge-podge of variables for high school students reached the .05 level of significance and 96% of them were significant at the 10^{-6} level. They involve, for example, such curious relations as between whether or not a teenager plays a musical instrument and to which of several Lutheran synods he belongs! With a sufficiently large sample, almost all of the 550 items of the MMPI are significant correlates of gender. In the early days of factor analysis employing orthogonal solutions, tests constructed with the aim of being factorially "pure" never turned out to be so. See other numerical examples in Meehl (1990a).

It has been suggested by defenders of null hypothesis testing that, although the crud factor may lead to a strong antecedent expectation of the falsity of H_0 for any given test, if tests are performed in several studies and each one yields significance in the predicted direction at some level α, we can combine these probabilities (the simplest way would be to merely multiply them, but I do not here enter the highly technical controversy about how best to put them together) and get a minuscule probability. There is something to this argument, but not very much. First, it still focuses attention upon the *p* value, that is, the improbability of getting a deviation if $\delta = 0$, and that is not the main point of the H_0-testing critics' objections. However, taking it as shown *in each individual experiment* that H_0 is false, we formulate the defender's argument thus: Assuming that everything is somewhat correlated with everything in the life sciences, the theorist has been making directional bets on the whole series of studies; if the atheoretical chance of guessing right, with the gigantic pots of theories and of observational variables that Meehl fantasizes, is one half for each guess (and neglecting imperfect power), then we should be allowed to use something like $.5^{10}$ when we look at 10 experiments that all "came out right." That is not a silly idea, and I readily concede that a big batch of H_0 refutations all going in the same direction must be deemed to provide considerable support for a theory that predicted them. However, I would not accept applying the multiplication principle of the probability calculus here, for two reasons.

First, it is an empirical fact over many domains of psychological research that we deal with a largely positive manifold. For example, in the case of intelligence and achievement, replicable negative correlations of subtests of an intelligence test, or of even "totally unrelated" achievement tests (e.g., spelling and art appreciation), almost never occur. Thorndike's dictum, which my undergraduate advisor Donald G. Paterson was fond of repeating, was that, "In psychology and sociology, all good things tend to go together." It is not, for instance, surprising to find a negative correlation between IQ and the incidence of tooth decay, although the explanation offered by dental hygienists in the 1920s was erroneous, because the true cause of this relationship is social class. So the chance of guessing the correct direction of a non-null difference is rarely only even, for a given empirical domain.

Second, most "interesting" psychological theories involve several "layers" or "levels" of theoretical constructs, and the layer that is closest in the causal chain to several observations may look pretty much the same in different theories. Taking my theory of schizotaxia as an example, one can easily come up with several plausible theories of why we might find a soft neurological sign (e.g., subclinical intention tremor) in the siblings of schizophrenes, theories that have no overlap in their "hard core" (Lakatos, 1970, 1974; Meehl, 1990a) with my schizotaxia conjecture, but that are eminently reasonable alternatives. There is no way to properly evaluate this point about accumulating successful directional predictions except to conduct large-scale archival studies as proposed by Faust and Meehl (1992).

The strongest rebuttal to this as a pro-H_0 argument is the well-known empirical finding in the research literature that you do not get highly consistent directional predictions for theories in most areas of soft psychology. You are lucky if your theory gives you successful directional predictions two thirds of the time. Here is where Sir Karl Popper enters the fray, in a way that does not depend on one's acceptance of his whole theory, but simply his emphasis upon the asymmetry between corroboration and falsification. It is incorrect to reason that if a substantive theory predicts a directional H_0 refutation correctly 7 times in 10 experiments and gets it wrong 3 times, then we have a "pretty good batting average for the theory." Given the crud factor, the seven correct predictions are weakly to moderately corroborative; but the three falsifications are—if accepted—*fatal*. It is not legitimate for the defender of H_0 testing to appeal to the problematic character of the other conjuncts in our corroboration formula to explain away the *modus tollens* falsifiers, meanwhile taking them for granted as unproblematic in considering the seven favorable outcomes. I repeat, this is not primarily a matter of α and β error rates; it goes through for any reasonable setting of the two kinds of errors.

STRONG VERSUS WEAK USE OF SIGNIFICANCE TESTS

In theory appraisal, significance tests may be used in a *strong* or a *weak* way, a distinction made by Meehl (1967; labeled as such in Meehl, 1990a, p. 116). This terminology is not a reference to the power function ($1-\beta$) or the size of α. It is not a question concerning the probability, confidence, confirmation, or corroboration of *H* at all. Rather, it asks what is the statistical hypothesis *H* that is being tested in order to appraise a substantive theory *T*, and how well can *H* serve to appraise *T*? The strong use of significance tests requires a strong theory, one capable of entailing a numerical value of the parameter, or a narrow range of theoretically tolerated values, or a specific function form (e.g., parabola) relating observables. Statistically significant deviation from the predicted point value, narrow interval, or curve type acts as a falsifier of the substantive theory. Hence, trying to refute the statistical hypothesis can function as a risky Popperian test or, for one who does not care for Popper but is an inductivist, as a Salmonian coincidence. The narrower the tolerated range of observable values, the riskier the test, and if the test is passed, the stronger the corroboration of the substantive theory. In the weak use, the theory is not powerful enough to make a point prediction or a narrow range prediction; it can say only that there is some nonzero correlation or some nonzero difference and, in almost all cases, to specify its algebraic direction. What makes this use weak, again, has nothing to do with the ratio $\alpha : \beta$, but involves the epistemic relation *between* the inference $H^*: \delta > 0$ and the alleged consequent confirmation of *T*.

The use of chi-square in social science provides a nice example of the difference between strong and weak uses. Pearson's (1900) classic article deriving what is often seen as his most important single contribution to biostatistics, the statistic chi-square, is titled: "On the Criterion That a Given System of Deviations From the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen From Random Sampling." Suppose some theory specifies a certain distribution of frequencies for various outcomes of throws of dice or colors of garden peas. The (strong) chi-square test is of whether the observed frequencies deviate from what the theory predicted. It was not until 4 years later that this same statistic was used in deriving the contingency coefficient, a statistic typically employed when the theory is not capable of predicting the tallies in various cells. The numerical prediction is based on multiplying marginals on the assumption of independence, but predicting (from a theory) that the two sets of categories being studied (such as eye color of siblings) are *not* statistically independent. In the strong use, the theory entails a set of numerical values, with a significant chi-square (indicating *departure* from the theoretically predicted values) functioning as a falsifier of the theory. In the weak use, the theory merely predicts "some (nonchance) relation," so a significant chi-square, falsifying statistical independence, tends to confirm the theory. What makes the weak use of a significance test weak is, of course, the crud factor, that there are numerous different theoretical ways that could account for some correlation or difference being nonzero; given the crud factor, a theory tested in this way takes only a small risk of being refuted if it is false. (Of course, the contingency coefficient could also be used strongly, if the substantive theory were strong enough to derive its numerical value, the *size* of a correlation of attributes. But in psychology that is almost never possible.)

There are special cases in the theory appraisal context in which refuting H_0 seems the appropriate statistical move rather than being merely a feeble substitute for something powerful (see Meehl, 1990a). The important one is the case of a strong theory according to which certain operations should yield a null effect. Here we have the strong use of a significance test that might be confused with the weak use. But here the point value predicted is $\delta = 0$, which the test attempts to refute, and *failing which refutation* the theory that predicts $\delta = 0$ is strongly corroborated. Recently, a radical theory in physics was proposed to the effect that gravitational force depended on something other than mass (the kind of elementary particle involved), or perhaps that there was a fifth force to be added to the received list (electromagnetic, gravitational, the strong and weak intranuclear forces). An experimental test involved lowering a plumbline into a deep mine shaft, for which the geological mineral surround was thoroughly known, and predicting its deflection, depending on the target element in the plumb. Result: A null effect. A good example in experimental psychology is latent learning, a disputed phenomenon with which my early research (with Kenneth MacCorquodale) was concerned. “Noncognitive” conditioning theories (Guthrie, Hull, Skinner) do not predict the latent learning phenomenon and cannot deal with it, although the Hullians attempted to do so by ingenious *ad-hoc*ery. If, during the “latent” (unreinforced) phase of such an experiment, rats show no increment in the strength of turning one way or the other in a T maze, but they show a preference when differential motivation or incentive is introduced, Tolman’s cognitive theory predicts evidence of something having been learned that did not manifest itself previously, whereas the three SR theories predict a null result. Although Tolman’s theory was too weak to predict a numerical point value, it did predict a deviation from the point $\delta = 0$, whereas that latter value was predicted by the three noncognitive competitor theories. Another example is telepathy, where the received theory—one might better say the received world view or metaphysic—predicts no transmission of information other than by photons, electrons, or sound waves. Thus, if a statistically significant amount of information is transmitted from one brain to another, absent any possibility of these kinds of physical pathways, we have evidence for a new kind of entity, force, or influence (which it is tempting to call “mental,” because there is a *cognitive content*; and, following Brentano, we take *intentionality* as the earmark of the “mental”).

The other helpful use of significant tests in the theoretical context is for a scientist deciding whether a certain line of research is worth pursuing. Suppose one of my graduate students has done an exploratory mini-experiment and tells me that in such and such a design 7 out of 12 rats made the predicted response. I agree that a description of a scientific experiment should be thought of rather like a recipe for baking cake (I believe it was B. F. Skinner who expressed this notion in conversation), so that what we ask is whether the recipe is described adequately enough that other competent cooks can make the cake by following it. If the experimental result replicates, we do not need a significance test. If it fails to replicate, the fact that it gave statistical significance when first reported is either dismissed as an improbable but always possible chance deviation, or simply left on the shelf as what the physicists call an “occult effect.” The reproducibility, smoothness, and highly characteristic features of cumulative records (drawn by the rat!) in Skinner’s 1938 classic are better science than any statistical significance tests, which is why the book contains none. But in some contexts matters are not so clear (e.g., maze learning—which is one of Skinner’s reasons for rejecting the maze as a research instrument). In the discovery phase, what I as a research scientist want to know to help make up my mind whether I should attempt to replicate a pilot study with a larger sample involves what degree of credence I give to it as a genuine effect on the basis of the 12 subjects run by the student. It seems rational to do a sign test on the 7:5 tally and to conclude that this is such weak evidence of a departure from "chance" as not to be worthwhile pursuing when I have other more plausible effects to explore. A 12:0 split ($p < .01$) might make it rational to try for a replication. This kind of decision, which is not rare in scientific work, would be impaired as to rationality if we did not have significance tests in our tool kit.

Both the strong and the weak uses have their characteristic dangers. The weak use, wrongly using significance tests as if refuting the null hypothesis gives powerful support to a weak theory, is obviously dangerous. In the strong use, refutation of the statistical hypothesis is a potential falsifier of the substantive theory that entailed it, and the degree of threat to the theory corresponds to the statistical power. However, falsification of the theory in practice always means falsification of the conjunction on the left of the conceptual formula for theory appraisal; and so one must be careful not to pessimistically apply one's confidence in the falsification of the conjunction to an equal confidence in the falsification of the theory, as explained earlier. (Saying T is false neglects 31 other ways the left side of the theory appraisal formula can be false.) But a second danger lurks here also, about which the statistician has nothing useful to tell us from the formalism. Having concluded that the theory T is literally false as it stands, and that this is correctly inferred (which we can be confident about depending on the smallness of α and the degree of certainty we attach to the other components of the appraisal formula), it might be reasoned that, therefore, the theory should be wholly discarded. Popper has been misread as saying this, but of course it does not directly follow from his position, especially once he began to emphasize the verisimilitude ("truthlikeness") concept. A theory can be literally false, meaning that the conjunction of all of its postulates is false, which means that at least one of them is false; so that the same argument about conjunctions and disjunctions that we applied previously to the appraisal formula applies also to the conjunction of postulates that make up the theory. Whether all scientific theories are literally false as commonly alleged I do not discuss, although I think it is erroneous; but assuming that were true, we still would want to appraise theories as to their verisimilitude. Unfortunately, the verisimilitude concept has thus far resisted rigorous explication. I agree with Popper that (pending such rigorous explication, which may not be possible in mathematical terms) we simply cannot dispense with it. Consider, for example, a question such as the accuracy of a certain newspaper story: The answer could vary from a minor and unimportant slip (such as getting somebody's middle initial wrong) to its being a totally fabricated account in a state-controlled propaganda sheet (such as Dr. Goebbels' fraudulent story of the Polish attack on the Gleiwitz radio transmitter, used as an excuse to invade Poland). I have made some tentative, rough-hewn steps toward formulating a metatheory of verisimilitude that takes a different approach from that of the logicians, but there is not space to discuss its merits and defects here (Meehl, 1990b, 1992a).

All theories in psychology, even in the most theoretically advanced "hard science" research domains, are at least incomplete, and in most domains, they are positively false as to what they do assert. Clarifying one's ideas about the strong and weak uses of significance tests is not as helpful as I had hoped when I first addressed this topic (Meehl, 1967). Answering a question about the point value or narrow interval of a statistical parameter does not tell us quite what we wanted to know in appraising the merits of a substantive theory. Even when the theory is so strong as to permit point predictions (as it is in special fields such as behavior genetics where, e.g., Meehl's dominant gene theory of schizotaxia leads to a point prediction about sibling concordance), the uncertainty of the auxiliaries, the doubtfulness of the *ceteris paribus* clause, the unreliability of measuring instruments, and so on, leave us wondering just what we should say when what appears to be a strong Popperian test is successfully passed or—even more so—is failed. "Error" here is a complicated mixture of conceptual error in the theory taken literally, along with conceptual error in the other conjuncts in the appraisal formula; the conceptual idealization leads to a numerical fuzziness, *even if* there were no errors of statistical inference of the kind conventionally addressed in statistics books. The history of science shows that—even for the most powerful of the exact sciences—numerical closeness to a theoretically predicted observational value is commonly taken as corroborative of a strong theory even if, strictly speaking, it is a falsifier because the observed value deviates "significantly" from the value predicted. That epistemological consideration greatly reduces the importance of the so-called "exact tests" or even "exact confidence intervals" dear to the hearts of statisticians and psychometricians.

A CORROBORATION INDEX FOR THEORY APPRAISAL

History of the advanced sciences shows that among the dozen or so properties and relations of theories that working scientists give weight to in theory appraisal, one of the most powerful is a theory's ability to derive numerical point predictions concerning the outcome of experimental measurements. Even when the numerical observation was known prior to a theory's formulation, if the derivation is "precise" (in a special sense I explain momentarily), it sometimes carries as much weight as the prediction of a numerically looser but novel or "surprising" observational finding. How often this happens, and under what social-cognitive circumstances, cannot be said until statistical studies of archival material in the history of science are conducted (Faust, 1984; Faust & Meehl, 1992; Meehl, 1992a, 1992b). In the meantime, one can only rely on anecdotal evidence as a working scientist, together with case studies by science historians. Brush (1989; see also Brush, 1995) has shown, for example, that the community of physicists was more impressed by general relativity's derivation of the anomalous Mercury perihelion—which had been known and puzzled over for a half-century before Einstein's 1916 exposition of general relativity—than they were by the allegedly "clinching" eclipse observations of 1919, so much emphasized by Eddington. I do not assert this ought to be the case; I merely point it out as the kind of historical episode that one must take into account in rational reconstruction of how scientists go about theory appraisal.

Independent convergence upon a theoretical number by qualitatively diverse experiments or statistical studies seems to exert an overwhelming influence on the convictions of the scientific community. A classic example is Perrin's (1916; see also Nye, 1972; Salmon, 1984) table showing the agreement of 13 independent observational methods for estimating Avagadro's number (the number of molecules in one gram molecular weight of a substance).

Perhaps the most powerful kind of theory corroboration occurs when a scientific theorist is able to combine *novelty* and *precision* as a derivation from theory, so that a new phenomenon or kind of qualitative observation is forecast from the new theory (an observation that had no rational expectation from background knowledge, from competing theories, or especially from the received theory), and the theory makes a numerically precise forecast that is confirmed. A theory's prediction of an observational numerical value—or if it is too weak to do that, its prediction that two or more different observational paths lead to a nearly identical inferred value—increases our strength of belief in the theory's verisimilitude, because if the theory had no truthlikeness, such a numerical agreement would be a Salmonian "strange coincidence." If one is a Popperian and disbelieves in induction (that is not my position, although I have strong Popperian sympathies in other respects), we say that the theory "surmounted a very risky hurdle"; it "took a big chance" in predicting with some exactitude something unknown. Failure to falsify an empirically dangerous test is what Popper meant by his term *corroboration*, as distinguished from the term *confirmation* employed by Carnap and other believers in induction. Fortunately, the dispute between inductivists and those philosophers of science who side with Popper in being skeptical about inductive logic need not concern us, because in this respect both lines of reasoning lead to the same methodological result.

But remember the fly in the ointment, arising from the logical structure of the corroboration formula presented earlier: the problematic character of the other conjuncts on the left side of the theory appraisal formula. It is often true in psychology that auxiliary theories A_x, instrumental theories A_i, or the *ceteris paribus* clause C_p are as problematic as the main theory T of interest. In addition to that problem, no psychologist believes that a theory is complete, although one may hope that none of the statements it makes are positively false. Also, in many areas of psychology it is usually pretentious to formulate strong theories that one cannot sincerely take literally. For these several reasons, precise predictions of numerical point values are difficult to come by. In rare cases (e.g., a dominant gene theory in behavior genetics) when we can make a numerical prediction, we cannot have much confidence that it will be exactly found in our observations because of the other components in the testing formula. This epistemologically loose situation leads us at times to count a "near miss" as almost as good as a numerical "hit," a practice that is also found in exact sciences such as astronomy, physics, and chemistry. Thus, what taken literally is a falsification is frequently an encouraging sign, especially in the early stages of theory construction, because "close enough" is a good reason for thinking that the theory has some merit (verisimilitude), despite not being exactly correct.

This reasoning leads us to seek a numerical index that will do justice to the methodological practice of giving positive weight to a "near miss" provided the prediction was sufficiently risky, given only background knowledge and absent our theory. The index I have proposed (Meehl, 1990a; 1990e) attempts to assess the riskiness of the observational prediction and the closeness to which it came. But I do not appraise closeness in terms of the conventional standard errors. What we mean by "closeness," as well as what we mean by "riskiness," depends on our background knowledge of the antecedently available range of numerical values, setting our theory (or competing theories) aside. I call that antecedent range of numerical values the *Spielraum*, following the 19th-century philosopher and statistician von Kries. The question of setting up a Spielraum and its rationale are discussed in Meehl (1990a), but the basic notion that risk and closeness should each be relativized to the Spielraum may be easily grasped in a common-sense, intuitive way:

> If I tell you that Meehl's theory of climate predicts that it will rain sometime next April, and this turns out to be the case, you will not be much impressed with my "predictive success." Nor will you be impressed if I predict more rain in April than in May, even showing three asterisks (for $p < .001$) in my t-test table! If I predict from my theory that it will rain on 7 of the 30 days in April, and it rains on exactly 7, you might perk up your ears a bit, but still you would be inclined to think of this as a "lucky coincidence." But suppose that I specify *which* 7 days in April it will rain and ring the bell; then you will start getting seriously interested in Meehl's meteorological conjectures. Finally, if I tell you that on April 4th it will rain 1.7 inches . . . and on April 9th 2.3 inches. . . and so forth, and get seven of these correct within reasonable tolerance, you will begin to think that Meehl's theory must

> have a lot going for it. You may believe that Meehl's theory of the weather, like all theories, is, when taken literally, false, since probably all theories are false in the eyes of God, but you will at least say, to use Popper's language, that it is beginning to look as if Meehl's theory has considerable *verisimilitude*, that is, "truth-likeness." (Meehl, 1978, pp. 817–818)

Or suppose I propound a genetic theory of mammalian embryology in reliance on which I claim to be able to predict the lengths of neonatal elephants' trunks with an average absolute error of .8 cm. You would not know whether to be impressed with my theory unless you knew the mean and (more important) the standard deviation of baby elephant trunks. Thus, if their mean length were 3 cm and the standard deviation 1 cm, my predictions average a 26% error and—worse—I could do just as well by simply guessing the mean each time.

To illustrate how a Spielraum might be specified when testing a theory in psychopathology, suppose I conjecture that 80%–90% of carefully diagnosed schizophrenes have become so on the basis of a schizotaxic brain, the latter due to an autosomal mutation completely penetrant for the integrative neural deficit. From this we infer that half of the siblings are schizotaxic. Reliability studies show that the *DSM* nosological label *schizophrenia* is about 90% correctly applied by skilled interviewers employing a quantified structured interview (e.g., SADS). Half the siblings should carry the dominant schizogene. Multiplying upper and lower bounds, we predict that from .360 to .405 of the sibs will be schizotaxic. This is the expected base rate range for a sibling schizotaxon estimated taxometrically (Meehl, 1995; Meehl & Yonce, 1994, 1996)—it does not assume perfect accuracy in classifying individuals—and it is the interval allowed by the theory, *I* in the index described shortly. If there were nothing heritable about schizophrenia, or the alleged schizotaxic soft neurology had nothing to do with it, the lower bound of a plausible Spielraum could be set at zero. An upper bound, as for DZ twins, would be .50, so here the Spielraum *S* is $.50 - .00 = .50$; hence the intolerance of the theory, *In* in the index, is $1 - I/S = 1 - (.405 - .360)/.50 = .91$, a pretty strong value for the "riskiness" component.

The same reasoning applies when a scientist appraises the accuracy of a measuring instrument or procedure. The standard error is the least informative in this respect, except in the strong use of a significance test. A percentage error is better, and it is commonly used among applied scientists (e.g., engineers). *Error in relation to an antecedently plausible range is often the most informative*. A numerical estimate with a 100-mile error is unacceptable in geography, mediocre in the solar system, brilliantly precise with respect to Alpha Centauri. In order to evaluate the bearing of an error (complement = "closeness," "accuracy," "precision"), what we most need to know is the range of values the physical world serves up to us, theory aside.

If this Spielraum reference is so important in powerful science appraisal, why do chemists and physicists not mention it? In those disciplines the notion is so universally understood and applied that they need no technical metatheoretical term for it. Everyone knows and daily applies the principle that a risky numerical prediction that turns out correct, or nearly correct, is a strong corroborator of the theory that accomplished this nice feat. Typically, it is in the late stages of refining and strongly challenging the limits of a well-entrenched theory that researchers in the exact sciences focus on the Popperian falsifiers, and at that point "close, but no cigar" becomes the ruling principle. Most of psychology is nowhere near that "ideal Popperian" stage of theory testing yet. Hence my former advocacy (Meehl, 1967, 1990a, 1990e) of the strong use of significance testing in psychology must be tempered.

Ideally, the theory is strong enough to derive a numerical observational value, such as the correlation of children's IQs with foster midparent IQs, or the incidence of a neurological sign in the siblings of schizophrenic patients on a dominant gene theory of schizotaxia (Meehl, 1962, 1989, 1990c, 1990d). In the life sciences, even a fairly strong theory cannot generate a numerical point value, but it may be strong enough to specify a numerical interval considerably narrower than the Spielraum. (Even a theoretical point value will be surrounded by a *measurement*-error interval, but I prefer to keep the *sampling* error dispersion out of my index.) The numerical interval that the theory allows being labeled *I* and the Spielraum *S*, we think of the ratio *I*/*S* as the theory's "relative tolerance" and the latter's complement $(1 - I/S)$ as the theory's intolerance, *In*. This functions as a crude measure of risk. Analogously, we measure the observed "closeness" as the complement of a relative error. If *D* is the deviation observed from the predicted point value, or the deviation from the edge of the interval predicted, then $(1 - D/S)$ is the experimental "closeness," *Cl*. Because I assume riskiness and closeness potentiate each other (i.e., they would, in a statistical study of the long-term history of successful and unsuccessful theories, display a Fisherian interaction effect), I multiply the two to get my confirmation index C_i provided by a particular experimental or statistical study. Thus, we have:

S: Spielraum;
I: interval tolerated by *T*;
I/*S*: relative tolerance of *T*;
In: $1 - (I/S)$ = intolerance of *T*;
D: deviation of observed value x_o from edge of tolerated interval (= error);
D/*S*: relative error;
Cl: $1 - (D/S)$ = closeness.

Then the corroboration index C_i for the particular experiment is defined as:

$$C_i = (Cl)(In)$$

Considering the best and worst cases, I standardize this index as C^*, which gives values lying in the familiar [0, 1] interval (Meehl, 1990a). I claim that the mean C^* value computed over a batch of empirical studies in which the theory is strong enough to make some sort of numerical prediction would be a more illuminating number to contemplate than a batch of significance tests refuting the null hypothesis.

One must resist the temptation to treat C^* as a probability merely because it lies in the interval [0, 1] as do its components. Even were they properly so interpretable, which is doubtful, no justification could be given for multiplying them in reliance on the independence assumption, clearly false. If the term *probable* is licit here, it is only in a loose epistemic (nonfrequentist, nonphysical) sense, that high values of C^* are not "likely" to arise in experiments whose numerical predictions flow from theories of low verisimilitude. I conceive this assumption as warranted mainly by history of science and scientific practice, although plausibility arguments can be offered in its favor from metatheoretical considerations. The (nonalgorithmic) epistemic move from a significance test or confidence interval—despite their being numerified—to a claim that the substantive theory T is thereby supported relies on exactly the same metatheoretical principle. But doesn't a quasi-Bayesian worry arise about the unknown distribution of values over the Spielraum, despite our eschewing a probability interpretation of C^*? (That we usually do not know this distribution is a further objection to writing $C^* = p$.) Suppose a crooked scientist routinely "predicts" numerical values at the middle of the Spielraum, setting a low-tolerance I, pretending these predictions are theoretically derived. Will this shady practice, relying on a sort of "least squares hopefulness," tend to produce a phony successful run of C^*s? No, it will not, for several reasons. First, most real-world experiments on interesting theories do not yield numerical values close to the Spielraum midpoint; if our pseudo-theorist sets the tolerance interval I low enough to be capable of generating an impressive C^* when the closeness component is good, the closeness component will not usually be good. Second, verisimilar competing theories that differentiate predicted locations will do better, as an algebraic necessity. Third, as the research data accumulate, it will become apparent that the mean and dispersion of our faker's errors are readily attributable to the average empirical bias of the midpoint prediction and the observed dispersion, a suspicious fact. (In the generalized correlation index, if the residual variance equals the predictand's total variance, C.I. = 0.) Fourth—the best safeguard—one does not treat a theorist's track record as if it arose from a black box predicting device. We require of any set of predictions to know how the *theory*, not the theorist, derives them. Guessing the center, even on the rare occasions when the center closely approximates the empirical mean, does not qualify conceptually as theory-derived. A mindless, automatic use of C^* as a truth-litmus test is, of course, unacceptable. *Rational interpretation of any statistic or index is subject to Carnap's Total Evidence Rule*.

Why propose an index of this sort, for which no exact "probabilifying" tables, such as those for F, chi-square, and t, exist or are likely to be constructed in the foreseeable future? Even assuming that those table values were precise for an actual life science situation (which they are not), their numbers denote conditional probabilities on the null hypothesis, and that is not the *kind* of probability that we seek in appraising a substantive scientific theory. I hope this has become clear from what has been explained here. I take the position (which I believe to be that of most philosophers of science and at least a sizable minority of applied statisticians and psychometricians) that it is better to have an index that appraises theory performance (corroboration, confirmation, degree of support) than it is to have a pseudo-exact number that only represents the first step in that chain of inference, namely, "something other than chance seems to be going on here." Accepting the view that no algorithm is capable of delivering a genuine probability$_1$ number for literal theoretical truth or quantifying a false theory's verisimilitude, we settle for less. My C^* is, of course, not a probability; it is an index of one aspect of empirical performance.

Second, I submit that the formula to some degree "speaks for itself," once its terms are understood. We have a standardized number lying in the familiar interval [0, 1] whose upper and lower bounds are approached when certain extreme epistemic situations have arisen. That is, if our substantive causal theory is so weak that it tolerates almost all of the antecedently possible numerical range (i.e., it takes a negligible risk), a "successful" observational outcome provides only weak support. If, on the other hand, the theory tolerates a moderate numerical range, or even a fairly small one, but the observations come nowhere near the predicted interval, then the theory surely receives no support from this experiment. Finally, if the theory is strong enough to prescribe a very narrow range of the antecedently conceivable Spielraum, and our experimental result falls within that range, we will get a corroboration index at the high end of the interval close to 1, which is obviously what we would want. That there are not precise confidence meanings attached to regions of this index should not bother anybody who understands the difference between falsification of the statistical hypothesis H_0 and resulting empirical corroboration of a substantive causal theory T, because nobody has ever proposed a useable algorithm for that important last step either.

Third, it is presently feasible to conduct Monte Carlo studies employing postulated theoretical structures with varying parameters. One can degrade the true theory *T* by increasingly severe deletions, additions, or amendments of its theoretical postulates, supplementing this impairment by throwing random measurement and sampling error on top of the increasingly deviant conceptual structures. In this way one can investigate the rough metrical properties of such an index.

Finally, there are plenty of scientific theories whose rise and fall is traceable in detail in the archives so that, given willingness to expend the effort, the empirical properties of the index in various research domains can be ascertained. It can be argued (Faust & Meehl, 1992) that calibration of such indexes on the basis of large-scale archival research in the history of science would be well worth the trouble, perhaps as valuable as the astronomers' gigantic catalog of stars or the recent mapping of the human genome. To develop this line of thought would require a long excursion into current controversies in metatheory (a preferable term to "philosophy of science," for reasons given in Meehl, 1992a, p. 340, fn. 2), but that is beyond the scope of this chapter. However, I urge the reader to recognize that the problem of null hypothesis testing as well as other problems of applied statistics depend at least as much on the resolution of controversies in metatheory as they do on clarifying our conceptual interpretations of the formalism. As I stated at the outset, hardly any of the controversies in the applications of mathematical statistics arise from strictly mathematical disagreements.

Why is meta-analysis—a powerful tool in social science research—not suitable for this purpose?

> Meta-analysis was developed to study outcomes of interventions (e.g., influence of class size, efficacy of psychotherapy or psychotropic drugs) rather than as a method of appraising the verisimilitude of substantive theories. We do not normally assume *theoretical corroboration* to be a monotone function, even stochastically, of *effect size*; and in developed sciences an observed value can, of course, be "too large" as often as "too small." . . . [Further, a] representative ("typical") effect size, whether of aggregated or disaggregated studies, is interpreted or qualified in meta-analysis via estimates of its standard error, emphasizing its trustworthiness as a numerical value. This statistical stability (under the laws of chance) is a very different question from how closely the effect approximates a theoretically predicted value. More importantly, it does not ask how "risky" the latter was in terms of the theoretically tolerated interval, in relation to the *a priori* range of possibilities. These two questions, taken jointly as the basis of all theoretical appraisal, require a different approach from that of evaluating technological outcomes in a pragmatic context. (Meehl, 1990e, p. 242)

The *C** index taps the interaction effect between risk and closeness, potentiating each by the other as a multiplier.

An objection to the proposed corroboration index C^* is an element of arbitrariness in specifying the Spielraum. I readily concede that this presents a problem, but I think not a serious one, for several nonexclusive reasons that tend cumulatively to minimize the problem while recognizing its existence, qualitatively speaking. First, in the "soft" areas of psychology, where refuting H_0 is most widely abused under the misconception that it provides a strong test of weak theories, the great preponderance of studies express their empirical findings with a dimensionless "pure" number, such as a probability or base rate lying in the interval [0, 1], or some form of correlation index (Pearson's r, biserial, tetrachoric, ϕ-coefficient, η, κ, and so on). These indexes of association are mathematically constructed so as to fall in the correlation interval [0, 1], although one must sometimes pay attention to additional constraints imposed by the data (e.g., the upper bound on a ϕ-coefficient set by disparity between the marginals). Factor loadings, standardized regression weights, and path coefficients (which are merely standardized regression weights for the subset of variables conjectured to be "causally upstream" from the predictand) are also pure numbers lying within the familiar [0, 1] interval. In noncorrelational studies employing randomized *experimental* treatments analyzed by Fisher's powerful method, the total sum of squares is decomposable into components associated with treatments and their interactions, with a residual that is a composite of sampling and measurement errors. The ratios of all these quantities to the total sum of squares lie in that standard interval [0, 1]. Although in a late stage of cliometric research it may be found that some fancier way of specifying the Spielraum is more informative, an obvious way to avoid arbitrariness in the present primitive state of actuarial metatheory is to define the Spielraum in terms of these pure number indexes—probabilities, correlations, loadings, β-weights, proportions of variance accounted for, and the like.

Second, in research domains where physical units of performance are the measure, one usually will have upper and lower bounds set by known dispositions of the organism (e.g., thresholds, physiological limits) or by the definition of the task. Thus in a maze-learning experiment, the worst score a rat can get is to enter every cul and the best is error-free performance. In the operant conditioning chamber, an animal may press the lever zero times in an hour or at a very high limiting rate (such as found in lean fixed-ratio schedules). These kinds of limits come not from the algebra of a correlation index, but from definitions, physical facts, and our "background knowledge" as the philosophers call it.

Third, complaining as I do against the atrocious practice of reporting significance tests without giving means and standard deviations, I, of course, would insist that presentation of C^* values be associated with statements of the tolerance interval derived from the theory and the Spielraum specified, permitting the reader to recompute a C^* using a Spielraum that seems more appropriate. That the probative weight of various values of C^* remains a matter of scholarly judgment is unavoidable, but it is no different from the present situation in which, having inferred from a refutation of H_0 that H^* holds (i.e., a directional difference $\delta > 0$ predicted by a weak theory T exists), one then uses one's (more or less rational) "scientific judgment" in appraising how much corroboration of the theory this statistical trend provides. I record my prediction that nothing short of a Faust–Meehl cliometric research program will do better than that.

That the values of C^* computed on the basis of two different Spielraum values are not linear transformations of one another over a range of experiments is regrettable. But I see no way to solve that; and here again, that is not very different from the familiar nonlinearity of different measures of association applied to a data set that is legitimately analyzable in more than one way. We do not ordinarily have occasion to compare corroboration of theories across domains. The working scientist's problem of appraisal is comparing two theories to explain the facts of a given domain of interest, or, if there is only one theory available, to ask whether it is well or weakly corroborated on the evidence to date. There are two important exceptions to this, however: the fund-granting agency deciding where to place its bets and the individual researcher planning the next few years of a research career. Significance tests obviously provide no decision algorithm for those cases either!

A researcher might adopt an inflated value of S to make it look good for a favored theory, or a very small one to make it look bad for a competing theory. Unfortunately, both components of the C^* index work in the same direction for such cases, but the reader can deflate the Spielraum if editors insist upon an adequate presentation of the data that go into the index. It is, of course, no objection to say that editors or referees *may* be careless in this respect, inasmuch as they are today routinely negligent in not requiring an adequate presentation of the data that enter into significance tests, or even confidence belts.

The tolerance I is less subject to this worry. Even a weak theory cannot plausibly be employed to derive a tolerance unless it is at least strong enough to justify the value of I employed, and here an inflated or deflated I works oppositely in the two components of the formula for C^*. If the investigator favors the theory and sets an inflated tolerance (so that the observed value is very likely to fall within the allowed range, which is itself therefore a large proportion of S), then the riskiness embodied in the other multiplier component is correspondingly reduced. In this respect, the assignment of a tolerance is analogous to the usual problem in significance testing of the trade-off between α and β errors.

CONCLUSIONS

Null-hypothesis testing has a proper place in social science research, but it has been criticized because of its widespread mindless abuse. Some have even proposed that H_0-testing be banned from American Psychological Association journals; such an editorial policy would constitute illegitimate censorship. Competent scholars persist in strong disagreement, ranging from some who think H_0-testing is pretty much all right as practiced, to others who think it is never appropriate. Most critics fall somewhere between these extremes, and they differ among themselves as to their main *reasons* for complaint. Under such circumstances, the proposed proscription of all H_0 testing would be a form of thought control, as foreign to the spirit of science as political correctness or religious orthodoxy. No such draconian measure should be contemplated when a few simple editorial rules would largely cure the disease, rules that editors should have been applying all along, given what is known. When null hypothesis testing is appropriate for the task:

1. Confidence intervals should be stated before H_0-tests are even mentioned.
2. Having stated a confidence interval, it would be permissible to add, "Because $\delta = \mu_1 - \mu_2 = 0$ falls outside the confidence belt, it is unplausible to explain the observed difference as arising from random sampling error." The misleading term *significant* is thus avoided; but it would be bizarre to require that a scientist state the confidence belt but forbid mention of a deductive consequence of it! I might be willing to forbid use of this cancerous term, if stating its legitimate *content* (nonmisleadingly) is allowed.
3. If inferences are made by contrasting significant and nonsignificant differences, the statistical powers must be stated.
4. A suitable measure of overlap must be computed (e.g., the percent of the experimental group exceeding the 10th, 25th, 50th, 75th, and 90th percentiles of the control group).
5. In the discussion section it must be explicitly stated that because the semantic contents of the counter-null hypothesis H^*: $\delta > 0$ and the substantive theory T are not equivalent—T implying H^*, but not conversely—proving H^* by refuting H_0 provides only weak corroboration of the theory. (This last is not imposing Meehl's philosophy of science on others; it is merely stating a truism found in any freshman logic text.)

Most important, researchers should distinguish statistical from epistemic questions, that is, when they are making an inference concerning a parameter (point value, range, slope, sign) from a statistic versus when they are appraising the verisimilitude of a substantive theory (causal, compositional, or historical) on the basis of the inferred parameters. The important distinction between these completely different kinds of inference should be clarified and emphasized in statistical texts and lectures. Failure to see this distinction engenders the tendency to think that, because α is chosen to be small ($p < .05$ or $< .01$), refuting H_0 somehow subjects the substantive theory T to a strong, "risky" test, so that passing this test provides strong evidentiary support for the theory, which it usually does not.

In general, the use of null hypothesis testing is more defensible in technological contexts than in theoretical contexts. Technological evaluation always involves at least an implicit weighting of utilities and disutilities of the two kinds of inferential errors. Sometimes the rational decision policy is betting on the reality of an observed algebraic sign whether statistically significant or not.

In correlational studies of (usually weak) theories in "soft" psychology, very few empirical variables are literally independent (i.e., H_0 is quasi-always false, theory aside), hence a theory-mediated prediction of H_0 being false usually can provide only weak theory corroboration. The crud factor ("everything tends to correlate with everything") corresponds in empirical research to the logician's warning that the third figure of the implicative syllogism is formally invalid. Given $p \supset q$ and q, inferring p is the formal fallacy of affirming the consequent. The substantive empirical meaning of this logical truism is the existence of numerous alternative theories capable of deriving a mere non-null difference.

The distinction between the strong and the weak use of significance tests is logical or epistemological; it is not a statistical issue. The weak use of significance tests asks merely whether the observations are attributable to "chance" (i.e., no relation exists) when a weak theory can only predict some sort of relation, but not what or how much. The strong use of significance tests asks whether observations differ significantly from the numerical values that a strong theory predicts, and it leads to the fourth figure of the syllogism—$p \supset q$, $\sim q$, infer $\sim p$—which is formally valid, the logician's *modus tollens* ("destroying mode"). Psychologists should work hard to formulate theories that, even if somewhat weak, permit derivation of numerical point values or narrow ranges, yielding the possibility of *modus tollens* refutations.

All psychological theories are imperfect, either incomplete or positively false as they are stated. Hence one expects confidence intervals (and H_0 testing as part of them) to falsify meritorious theories that possess enough verisimilitude to deserve continued amendment, testing, and expansion. In the early stages of a good theory's life, a "near-miss" of prediction is an encouraging finding despite being a literal falsification of the theory and associated conjectures involved in the observational test. It is therefore desirable to have a quantitative index of how risky the test was and how close the prediction came, and for this purpose I have proposed a corroboration index C^* that potentiates closeness by risk.

Finally, the most important property of an empirical finding is intersubjective replicability, that other investigators, relying on the description of what was done, will (almost always) make the same (or closely similar) observations.

REFERENCES

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

Brush, S. G. (1989). Prediction and theory evaluation: the case of light bending. *Science, 246,* 1124–1129.

Brush, S. G. (1995). Dynamics of theory change: The role of predictions. *PSA 1994, 2,* 133–145.

Carnap, R. (1945). The two concepts of probability. *Philosophy and Phenomenological Research, 5,* 513–532. Reprinted in H. Feigl & W. Sellars (Eds.), *Readings in philosophical analysis* (pp. 330–348). New York: Appleton–Century–Crofts, 1949. Reprinted in H. Feigl & M. Broadbeck (Eds.), *Readings in the philosophy of science* (pp. 438–455). New York: Appleton–Century–Crofts, 1953.

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Earman, J. (1986). *A primer on determinism*. Boston: D. Reidel.

Faust, D. (1984). *The limits of scientific reasoning*. Minneapolis: University of Minnesota Press.

Faust, D., & Meehl, P. E. (1992). Using scientific methods to resolve enduring questions within the history and philosophy of science: Some illustrations. *Behavior Therapy, 23,* 195–211.

Hogben, L. (1968). *Statistical theory.* New York: Norton.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds*.), Criticism and the growth of knowledge* (pp. 91–195). Cambridge, England: Cambridge University Press. Reprinted in J. Worrall & G. Currie (Eds.), *Imre Lakatos: Philosophical papers: Vol. I. The methodology of scientific research programmes* (pp. 8–101). New York: Cambridge University Press, 1978.

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

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

Meehl, P. E. (1962). Schizotaxia, schizotypy, schizophrenia*. American Psychologist, 17,* 827–838. Reprinted in Meehl, *Psychodiagnosis: Selected papers* (pp. 135–155). Minneapolis: University of Minnesota Press, 1973.

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

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

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

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

Meehl, P. E. (1990b*). Corroboration and verisimilitude: Against Lakatos'* "*sheer leap of faith*" (Working Paper No. MCPS–90–01). Minneapolis: University of Minnesota, Center for Philosophy of Science.

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

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

Meehl, P. E. (1990e). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244. Also in R. E. Snow & D. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 13–59). Hillsdale, NJ: Erlbaum, 1991.

Meehl, P. E. (1992a). Cliometric metatheory: The actuarial approach to empirical, history-based philosophy of science. *Psychological Reports, 71,* 339–467.

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

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist, 50,* 266–275.

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

Meehl, P. E., & Yonce, L. J. (1996). Taxometric analysis: II. Detecting taxonicity using covariance of two quantitative indicators in successive intervals of a third indicator (MAXCOV procedure). *Psychological Reports, 78,* 1091–1227.

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

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

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

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series V.1,* 157–175.

Perrin, J. B. (1916). *Atoms* (D. L. Hammick, Trans.). New York: Van Nostrand.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books. (Original work published 1934)

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

Popper, K. R. (1983). *Postscript to the logic of scientific discovery: Vol. I. Realism and the aim of science*. Totowa, NJ: Rowman and Littlefield.

Reichenbach, H. (1938). *Experience and prediction*. Chicago: University of Chicago Press.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

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

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

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Simon, H. A. (1945). Statistical tests as a basis for "yes-no" choices. *Journal of the American Statistical Association, 40,* 80–84.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton–Century.