The Earth Is Round ($p < .05$)

Jacob Cohen

After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including its near-universal misinterpretation of p *as the probability that* H_0 *is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects* H_0 *one thereby affirms the theory that led to the test. Exploratory data analysis and the use of graphic methods, a steady improvement in and a movement toward standardization in measurement, an emphasis on estimating effect sizes using confidence intervals, and the informed use of available statistical methods is suggested. For generalization, psychologists must finally rely, as has been done in all the older sciences, on replication.*

I make no pretense of the originality of my remarks in this article. One of the few things we, as psychologists, have learned from over a century of scientific study is that at age three score and 10, originality is not to be expected. David Bakan said back in 1966 that his claim that "a great deal of mischief has been associated" with the test of significance "is hardly original," that it is "what 'everybody knows,'" and that "to say it 'out loud' is . . . to assume the role of the child who pointed out that the emperor was really outfitted in his underwear" (p. 423). If it was hardly original in 1966, it can hardly be original now. Yet this naked emperor has been shamelessly running around for a long time.

Like many men my age, I mostly grouse. My harangue today is on testing for statistical significance, about which Bill Rozeboom (1960) wrote 33 years ago, "The statistical folkways of a more primitive past continue to dominate the local scene" (p. 417).

And today, they continue to continue. And we, as teachers, consultants, authors, and otherwise perpetrators of quantitative methods, are responsible for the ritualization of null hypothesis significance testing (NHST; I resisted the temptation to call it statistical hypothesis inference testing) to the point of meaninglessness and beyond. I argue herein that NHST has not only failed to support the advance of psychology as a science but also has seriously impeded it.

Consider the following: A colleague approaches me with a statistical problem. He believes that a generally rare disease does not exist at all in a given population, hence H_0: $P = 0$. He draws a more or less random sample of 30 cases from this population and finds that one of the cases has the disease, hence $P_s = 1/30 = .033$. He is not sure how to test H_0, chi-square with Yates's (1951) correction or the Fisher exact test, and wonders whether he has enough power. Would you believe it? And would you believe that if he tried to publish this result without a significance test, one or more reviewers might complain? It could happen.

Almost a quarter of a century ago, a couple of sociologists, D. E. Morrison and R. E. Henkel (1970), edited a book entitled *The Significance Test Controversy*. Among the contributors were Bill Rozeboom (1960), Paul Meehl (1967), David Bakan (1966), and David Lykken (1968). Without exception, they damned NHST. For example, Meehl described NHST as "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" (p. 265). They were, however, by no means the first to do so. Joseph Berkson attacked NHST in 1938, even before it sank its deep roots in psychology. Lancelot Hogben's book-length critique appeared in 1957. When I read it then, I was appalled by its rank apostasy. I was at that time well trained in the current Fisherian dogma and had not yet heard of Neyman-Pearson (try to find a reference to them in the statistics texts of that day—McNemar, Edwards, Guilford, Walker). Indeed, I had already had some dizzying success as a purveyor of plain and fancy NHST to my fellow clinicians in the Veterans Administration.

What's wrong with NHST? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is "Given these data, what is the probability that H_0 is true?" But as most of us know, what it tells us is "Given that H_0 is true, what is the probability of these (or more extreme) data?" These are not the same, as has been pointed out many times over the years by the contributors to the Morrison–Henkel (1970) book, among

J. Bruce Overmier served as action editor for this article.

This article was originally an address given for the Saul B. Sells Memorial Lifetime Achievement Award, Society of Multivariate Experimental Psychology, San Pedro, California, October 29, 1993.

I have made good use of the comments made on a preliminary draft of this article by Patricia Cohen and other colleagues: Robert P. Abelson, David Bakan, Michael Borenstein, Robyn M. Dawes, Ruma Falk, Gerd Gigerenzer, Charles Greenbaum, Raymond A. Katzell, Donald F. Klein, Robert S. Lee, Paul E. Meehl, Stanley A. Mulaik, Robert Rosenthal, William W. Rozeboom, Elia Sinaiko, Judith D. Singer, and Bruce Thompson. I also acknowledge the help I received from reviewers David Lykken, Matt McGue, and Paul Slovic.

Correspondence concerning this article should be addressed to Jacob Cohen, Department of Psychology, New York University, 6 Washington Place, 5th Floor, New York, NY 10003.

 0003-066X/94/$2.00
Vol. 49, No. 12, 997-1003

others, and, more recently and emphatically, by Meehl (1978, 1986, 1990a, 1990b), Gigerenzer (1993), Falk and Greenbaum (in press), and yours truly (Cohen, 1990).

The Permanent Illusion

One problem arises from a misapplication of deductive syllogistic reasoning. Falk and Greenbaum (in press) called this the "illusion of probabilistic proof by contradiction" or the "illusion of attaining improbability." Gigerenzer (1993) called it the "permanent illusion" and the "Bayesian Id's wishful thinking," part of the "hybrid logic" of contemporary statistical inference—a mishmash of Fisher and Neyman–Pearson, with invalid Bayesian interpretation. It is the widespread belief that the level of significance at which H_0 is rejected, say .05, is the probability that it is correct or, at the very least, that it is of low probability.

The following is almost but not quite the reasoning of null hypothesis rejection:

If the null hypothesis is correct, then this datum (D) can not occur.
It has, however, occurred.
Therefore, the null hypothesis is false.

If this were the reasoning of H_0 testing, then it would be formally correct. It would be what Aristotle called the modus tollens, denying the antecedent by denying the consequent. But this is not the reasoning of NHST. Instead, it makes this reasoning probabilistic, as follows:

If the null hypothesis is correct, then these data are highly unlikely.
These data have occurred.
Therefore, the null hypothesis is highly unlikely.

By making it probabilistic, it becomes invalid. Why? Well, consider this:

The following syllogism is sensible and also the formally correct modus tollens:

If a person is a Martian, then he is not a member of Congress.
This person is a member of Congress.
Therefore, he is not a Martian.

Sounds reasonable, no? This next syllogism is not sensible because the major premise is wrong, but the reasoning is as before and still a formally correct modus tollens:

If a person is an American, then he is not a member of Congress. (WRONG!)
This person is a member of Congress.
Therefore, he is not an American.

If the major premise is made sensible by making it probabilistic, not absolute, the syllogism becomes formally incorrect and leads to a conclusion that is not sensible:

If a person is an American, then he is probably not a member of Congress. (TRUE, RIGHT?)
This person is a member of Congress.
Therefore, he is probably not an American. (Pollard & Richardson, 1987)

This is formally exactly the same as

If H_0 is true, then this result (statistical significance) would probably not occur.
This result has occurred.
Then H_0 is probably not true and therefore formally invalid.

This formulation appears at least implicitly in article after article in psychological journals and explicitly in some statistics textbooks—"the illusion of attaining improbability."

Why $P(D|H_0) \neq P(H_0|D)$

When one tests H_0, one is finding the probability that the data (D) could have arisen if H_0 were true, $P(D|H_0)$. If that probability is small, then it can be concluded that if H_0 is true, then D is unlikely. Now, what really is at issue, what is always the real issue, is the probability that H_0 is true, given the data, $P(H_0|D)$, the inverse probability. When one rejects H_0, one wants to conclude that H_0 is unlikely, say, $p < .01$. The very reason the statistical test is done is to be able to reject H_0 because of its unlikelihood! But that is the posterior probability, available only through Bayes's theorem, for which one needs to know $P(H_0)$, the probability of the null hypothesis before the experiment, the "prior" probability.

Now, one does not normally know the probability of H_0. Bayesian statisticians cope with this problem by positing a prior probability or distribution of probabilities. But an example from psychiatric diagnosis in which one knows $P(H_0)$ is illuminating:

The incidence of schizophrenia in adults is about 2%. A proposed screening test is estimated to have at least 95% accuracy in making the positive diagnosis (sensitivity) and about 97% accuracy in declaring normality (specificity). Formally stated, $P(\text{normal}|H_0) \approx .97$, $P(\text{schizophrenia}|H_1) > .95$. So, let

H_0 = The case is normal, so that
H_1 = The case is schizophrenic, and
D = The test result (the data) is positive for schizophrenia.

With a positive test for schizophrenia at hand, given the more than .95 assumed accuracy of the test, $P(D|H_0)$—the probability of a positive test given that the case is normal—is less than .05, that is, significant at $p < .05$. One would reject the hypothesis that the case is normal and conclude that the case has schizophrenia, as it happens mistakenly, but within the .05 alpha error. But that's not the point.

The probability of the case being normal, $P(H_0)$, given a positive test (D), *that is*, $P(H_0|D)$, is not what has just been discovered however much it sounds like it and however much it is wished to be. It is not true that the probability that the case is normal is less than .05, nor is it even unlikely that it is a normal case. By a Bayesian maneuver, this inverse probability, the probability that

the case is normal, given a positive test for schizophrenia, is about .60! The arithmetic follows:

$$P(H_0 \mid D)$$

$$= \frac{P(H_0) * P(\text{test wrong} \mid H_0)}{P(H_0) * P(\text{test wrong} \mid H_0) + P(H_1) * P(\text{test correct} \mid H_1)}$$

$$= \frac{(.98)(.03)}{(.98)(.03) + (.02)(.95)} = \frac{.0294}{.0294 + .0190} = .607$$

The situation may be made clearer by expressing it approximately as a 2 × 2 table for 1,000 cases. The case actually is

Result	Normal	Schiz	Total
Negative test (Normal)	949	1	950
Positive test (Schiz)	30	20	50
Total	979	21	1,000

As the table shows, the conditional probability of a normal case for those testing as schizophrenic is not small—of the 50 cases testing as schizophrenics, 30 are false positives, actually normal, 60% of them!

This extreme result occurs because of the low base rate for schizophrenia, but it demonstrates how wrong one can be by considering the p value from a typical significance test as bearing on the truth of the null hypothesis for a set of data.

It should not be inferred from this example that all null hypothesis testing requires a Bayesian prior. There is a form of H_0 testing that has been used in astronomy and physics for centuries, what Meehl (1967) called the "strong" form, as advocated by Karl Popper (1959). Popper proposed that a scientific theory be tested by attempts to falsify it. In null hypothesis testing terms, one takes a central prediction of the theory, say, a point value of some crucial variable, sets it up as the H_0, and challenges the theory by attempting to reject it. This is certainly a valid procedure, potentially even more useful when used in confidence interval form. What I and my ilk decry is the "weak" form in which theories are "confirmed" by rejecting null hypotheses.

The inverse probability error in interpreting H_0 is not reserved for the great unwashed, but appears many times in statistical textbooks (although frequently together with the correct interpretation, whose authors apparently think they are interchangeable). Among the distinguished authors making this error are Guilford, Nunnally, Anastasi, Ferguson, and Lindquist. Many examples of this error are given by Robyn Dawes (1988, pp. 70–75); Falk and Greenbaum (in press); Gigerenzer (1993, pp. 316–329), who also nailed R. A. Fisher (who emphatically rejected Bayesian theory of inverse probability but slipped into invalid Bayesian interpretations of NHST (p. 318); and Oakes (1986, pp. 17–20), who also nailed me for this error (p. 20).

The illusion of attaining improbability or the Bayesian Id's wishful thinking error in using NHST is very easy to make. It was made by 68 out of 70 academic psychologists studied by Oakes (1986, pp. 79–82). Oakes incidentally offered an explanation of the neglect of power analysis because of the near universality of this inverse probability error:

> After all, why worry about the probability of obtaining data that will lead to the rejection of the null hypothesis if it is false when your analysis gives you the actual probability of the null hypothesis being false? (p. 83)

A problem that follows readily from the Bayesian Id's wishful thinking error is the belief that after a successful rejection of H_0, it is highly probable that replications of the research will also result in H_0 rejection. In their classic article "The Belief in the Law of Small Numbers," Tversky and Kahneman (1971) showed that because people's intuitions that data drawn randomly from a population are highly representative, most members of the audience at an American Psychological Association meeting and at a mathematical psychology conference believed that a study with a significant result would replicate with a significant result in a small sample (p. 105). Of Oakes's (1986) academic psychologists 42 out of 70 believed that a t of 2.7, with $df = 18$ and $p = .01$, meant that if the experiment were repeated many times, a significant result would be obtained 99% of the time. Rosenthal (1993) said with regard to this replication fallacy that "Nothing could be further from the truth" (p. 542f) and pointed out that given the typical .50 level of power for medium effect sizes at which most behavioral scientists work (Cohen, 1962), the chances are that in three replications only one in eight would result in significant results, in all three replications, and in five replications, the chance of as many as three of them being significant is only 50:50.

An error in elementary logic made frequently by NHST proponents and pointed out by its critics is the thoughtless, usually implicit, conclusion that if H_0 is rejected, then the theory is established: If A then B; B therefore A. But even the valid form of the syllogism (if A then B; not B therefore not A) can be misinterpreted. Meehl (1990a, 1990b) pointed out that in addition to the theory that led to the test, there are usually several auxiliary theories or assumptions and ceteris paribus clauses and that it is the logical product of these that is counterpoised against H_0. Thus, when H_0 is rejected, it can be because of the falsity of any of the auxiliary theories about instrumentation or the nature of the psyche or of the ceteris paribus clauses, and not of the substantive theory that precipitated the research.

So even when used and interpreted "properly," with a significance criterion (almost always $p < .05$) set a priori (or more frequently understood), H_0 has little to commend it in the testing of psychological theories in its usual reject-H_0-confirm-the-theory form. The ritual dichotomous reject–accept decision, however objective and administratively convenient, is not the way any science is done. As Bill Rozeboom wrote in 1960, "The primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the de-

gree to which one . . . believes the hypothesis . . . being tested" (p. 420)

The Nil Hypothesis

Thus far, I have been considering H_0s in their most general sense—as propositions about the state of affairs in a population, more particularly, as some specified value of a population parameter. Thus, "the population mean difference is 4" may be an H_0, as may be "the proportion of males in this population is .75" and "the correlation in this population is .20." But as almost universally used, the null in H_0 is taken to mean nil, zero. For Fisher, the null hypothesis was the hypothesis to be nullified. As if things were not bad enough in the interpretation, or misinterpretation, of NHST in this general sense, things get downright ridiculous when H_0 is to the effect that the effect size (ES) is 0—that the population mean difference is 0, that the correlation is 0, that the proportion of males is .50, that the raters' reliability is 0 (an H_0 that can almost always be rejected, even with a small sample—Heaven help us!). Most of the criticism of NHST in the literature has been for this special case where its use may be valid only for true experiments involving randomization (e.g., controlled clinical trials) or when any departure from pure chance is meaningful (as in laboratory experiments on clairvoyance), but even in these cases, confidence intervals provide more information. I henceforth refer to the H_0 that an ES = 0 as the "nil hypothesis."

My work in power analysis led me to realize that the nil hypothesis is always false. If I may unblushingly quote myself,

> It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null hypothesis is always false, what's the big deal about rejecting it? (p. 1308)

I wrote that in 1990. More recently I discovered that in 1938, Berkson wrote

> It would be agreed by statisticians that a large sample is always better than a small sample. If, then, we know in advance the P that will result from an application of the Chi-square test to a large sample, there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all. (p. 526f)

Tukey (1991) wrote that "It is foolish to ask 'Are the effects of A and B different?' They are always different—for some decimal place" (p. 100).

The point is made piercingly by Thompson (1992):

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data on hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they are tired. This tautology has created considerable damage as regards the cumulation of knowledge. (p. 436)

In an unpublished study, Meehl and Lykken cross-tabulated 15 items for a sample of 57,000 Minnesota high school students, including father's occupation, father's education, mother's education, number of siblings, sex, birth order, educational plans, family attitudes toward college, whether they liked school, college choice, occupational plan in 10 years, religious preference, leisure time activities, and high school organizations. All of the 105 chi-squares that these 15 items produced by the cross-tabulations were statistically significant, and 96% of them at $p < .000001$ (Meehl, 1990b).

One might say, "With 57,000 cases, relationships as small as a Cramer ϕ of .02–.03 will be significant at $p < .000001$, so what's the big deal?" Well, the big deal is that many of the relationships were much larger than .03. Enter the Meehl "crud factor," more genteelly called by Lykken "the ambient correlation noise." In soft psychology, "Everything is related to everything else." Meehl acknowledged (1990b) that neither he nor anyone else has accurate knowledge about the size of the crud factor in a given research domain, *"but the notion that the correlation between arbitrarily paired trait variables will be, while not literally zero, of such minuscule size as to be of no importance, is surely wrong"* (p. 212, italics in original).

Meehl (1986) considered a typical review article on the evidence for some theory based on nil hypothesis testing that reports a 16:4 box score in favor of the theory. After taking into account the operation of the crud factor, the bias against reporting and publishing "negative" results (Rosenthal's, 1979, "file drawer" problem), and assuming power of .75, he estimated the likelihood ratio of the theory against the crud factor as 1:1. Then, assuming that the prior probability of theories in soft psychology is $\leq .10$, he concluded that the Bayesian posterior probability is also $\leq .10$ (p. 327f). So a 16:4 box score for a theory becomes, more realistically, a 9:1 odds ratio against it.

Meta-analysis, with its emphasis on effect sizes, is a bright spot in the contemporary scene. One of its major contributors and proponents, Frank Schmidt (1992), provided an interesting perspective on the consequences of current NHST-driven research in the behavioral sciences. He reminded researchers that, given the fact that the nil hypothesis is always false, the rate of Type I errors is 0%, not 5%, and that only Type II errors can be made, which run typically at about 50% (Cohen, 1962; Sedlmeier & Gigerenzer, 1989). He showed that typically, the sample effect size necessary for significance is notably larger than the actual population effect size and that the average of the statistically significant effect sizes is much larger than the actual effect size. The result is that people who do focus on effect sizes end up with a substantial positive bias in their effect size estimation. Furthermore, there is the irony that the "sophisticates" who use procedures to adjust their alpha error for multiple tests (using Bonferroni, Newman-Keuls, etc.) are adjusting for a nonexistent alpha error, thus reduce their power, and, if lucky enough to get a significant result, only end up grossly overestimating the population effect size!

Because NHST p values have become the coin of the realm in much of psychology, they have served to

inhibit its development as a science. Go build a quantitative science with p values! All psychologists know that *statistically significant* does not mean plain-English significant, but if one reads the literature, one often discovers that a finding reported in the Results section studded with asterisks implicitly becomes in the Discussion section highly significant or very highly significant, important, big!

Even a correct interpretation of p values does not achieve very much, and has not for a long time. Tukey (1991) warned that if researchers fail to reject a nil hypothesis about the difference between A and B, all they can say is that the direction of the difference is "uncertain." If researchers reject the nil hypothesis then they can say they can be pretty sure of the direction, for example, "A is larger than B." But if all we, as psychologists, learn from a research is that A is larger than B ($p < .01$), we have not learned very much. And this is typically all we learn. Confidence intervals are rarely to be seen in our publications. In another article (Tukey, 1969), he chided psychologists and other life and behavior scientists with the admonition "Amount, as well as direction is vital" and went on to say the following:

> The physical scientists have learned much by storing up amounts, not just directions. If, for example, elasticity had been confined to "When you pull on it, it gets longer!," Hooke's law, the elastic limit, plasticity, and many other important topics could not have appeared (p. 86). . . . Measuring the right things on a communicable scale lets us stockpile information about amounts. Such information can be useful, whether or not the chosen scale is an interval scale. Before the second law of thermodynamics—and there were many decades of progress in physics and chemistry before it appeared—the scale of temperature was not, in any nontrivial sense, an interval scale. Yet these decades of progress would have been impossible had physicists and chemists refused either to record temperatures or to calculate with them. (p. 80)

In the same vein, Tukey (1969) complained about correlation coefficients, quoting his teacher, Charles Winsor, as saying that they are a dangerous symptom. Unlike regression coefficients, correlations are subject to vary with selection as researchers change populations. He attributed researchers' preference for correlations to their avoidance of thinking about the units with which they measure.

> Given two perfectly meaningless variables, one is reminded of their meaninglessness when a regression coefficient is given, since one wonders how to interpret its value. . . . Being so uninterested in our variables that we do not care about their units can hardly be desirable. (p. 89)

The major problem with correlations applied to research data is that they can not provide useful information on causal strength because they change with the degree of variability of the variables they relate. Causality operates on single instances, not on populations whose members vary. The effect of A on B for me can hardly depend on whether I'm in a group that varies greatly in A or another that does not vary at all. It is not an accident that causal modeling proceeds with regression and not correlation coefficients. In the same vein, I should note that standardized effect size measures, such as d and f, developed in power analysis (Cohen, 1988) are, like correlations, also dependent on population variability of the dependent variable and are properly used only when that fact is kept in mind .

To work constructively with "raw" regression coefficients and confidence intervals, psychologists have to start respecting the units they work with, or develop measurement units they can respect enough so that researchers in a given field or subfield can agree to use them. In this way, there can be hope that researchers' knowledge can be cumulative. There are few such in soft psychology. A beginning in this direction comes from meta-analysis, which, whatever else it may accomplish, has at least focused attention on effect sizes. But imagine how much more fruitful the typical meta-analysis would be if the research covered used the same measures for the constructs they studied. Researchers could get beyond using a mass of studies to demonstrate convincingly that "if you pull on it, it gets longer."

Recall my example of the highly significant correlation between height and intelligence in 14,000 school children that translated into a regression coefficient that meant that to raise a child's IQ from 100 to 130 would require giving enough growth hormone to raise his or her height by 14 feet (Cohen, 1990).

What to Do?

First, don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist.

Second, even before we, as psychologists, seek to generalize from our data, we must seek to understand and improve them. A major breakthrough to the approach to data, emphasizing "detective work" rather than "sanctification" was heralded by John Tukey in his article "The Future of Data Analysis" (1962) and detailed in his seminal book *Exploratory Data Analysis* (EDA; 1977). EDA seeks not to vault to generalization to the population but by simple, flexible, informal, and largely graphic techniques aims for understanding the set of data in hand. Important contributions to graphic data analysis have since been made by Tufte (1983, 1990), Cleveland (1993; Cleveland & McGill, 1988), and others. An excellent chapter-length treatment by Wainer and Thissen (1981), recently updated (Wainer & Thissen, 1993), provides many useful references, and statistical program packages provide the necessary software (see, for an example, Lee Wilkinson's [1990] SYGRAPH, which is presently being updated).

Forty-two years ago, Frank Yates, a close colleague and friend of R. A. Fisher, wrote about Fisher's "Statistical Methods for Research Workers" (1925/1951),

> It has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data . . . and too little to the estimates of the magnitude of the effects they are estimating (p. 32).

Thus, my third recommendation is that, as researchers, we routinely report effect sizes in the form of confidence limits. "Everyone knows" that confidence intervals contain all the information to be found in significance tests and much more. They not only reveal the status of the trivial nil hypothesis but also about the status of non-nil null hypotheses and thus help remind researchers about the possible operation of the crud factor. Yet they are rarely to be found in the literature. I suspect that the main reason they are not reported is that they are so embarrassingly large! But their sheer size should move us toward improving our measurement by seeking to reduce the unreliable and invalid part of the variance in our measures (as Student himself recommended almost a century ago). Also, their width provides us with the analogue of power analysis in significance testing—larger sample sizes reduce the size of confidence intervals as they increase the statistical power of NHST. A new program covers confidence intervals for mean differences, correlation, cross-tabulations (including odds ratios and relative risks), and survival analysis (Borenstein, Cohen, & Rothstein, in press). It also produces Birnbaum's (1961) "confidence curves," from which can be read all confidence intervals from 50% to 100%, thus obviating the necessity of choosing a specific confidence level for presentation.

As researchers, we have a considerable array of statistical techniques that can help us find our way to theories of some depth, but they must be used sensibly and be heavily informed by informed judgment. Even null hypothesis testing complete with power analysis can be useful if we abandon the rejection of point nil hypotheses and use instead "good-enough" range null hypotheses (e.g., "the effect size is no larger than 8 raw score units, or $d = .5$), as Serlin and Lapsley (1993) have described in detail. As our measurement and theories improve, we can begin to achieve the Popperian principle of representing our theories as null hypotheses and subjecting them to challenge, as Meehl (1967) argued many years ago. With more evolved psychological theories, we can also find use for likelihood ratios and Bayesian methods (Goodman, 1993; Greenwald, 1975). We quantitative behavioral scientists need not go out of business.

Induction has long been a problem in the philosophy of science. Meehl (1990a) attributed to the distinguished philosopher Morris Raphael Cohen the saying "All logic texts are divided into two parts. In the first part, on deductive logic, the fallacies are explained; in the second part, on inductive logic, they are committed" (p. 110). We appeal to inductive logic to move from the particular results in hand to a theoretically useful generalization. As I have noted, we have a body of statistical techniques, that, used intelligently, can facilitate our efforts. But given the problems of statistical induction, we must finally rely, as have the older sciences, on replication.

REFERENCES

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 1–29.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33,* 526–542.

Birnbaum, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Statistical Association, 56,* 246–249.

Borenstein, M., Cohen, J., & Rothstein, H. (in press). *Confidence intervals, effect size, and power* [Computer program]. Hillsdale, NJ: Erlbaum.

Cleveland, W. S. (1993). *Visualizing data.* Summit, NJ: Hobart.

Cleveland, W. S., & McGill, M. E. (Eds.). (1988). *Dynamic graphics for statistics.* Belmont, CA: Wadsworth.

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 69,* 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Dawes, R. M. (1988). *Rational choice in an uncertain world.* San Diego, CA: Harcourt Brace Jovanovich.

Falk, R., & Greenbaum, C. W. (in press). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology.*

Fisher, R. A. (1951). *Statistical methods for research workers.* Edinburgh, Scotland: Oliver & Boyd. (Original work published 1925)

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Ed.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Goodman, S. N. (1993). *P* values, hypothesis tests, and likelihood implications for epidemiology: Implications of a neglected historical debate. *American Journal of Epidemiology, 137,* 485–496.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Hogben, L. (1957). *Statistical theory.* London: Allen & Unwin.

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

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

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. (1986). What social scientists don't understand. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science: Pluralisms and subjectivities* (pp. 315–338). Chicago: University of Chicago Press.

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

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*(Monograph Suppl. 1–V66), 195–244.

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

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: Wiley.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 102,* 159–163.

Popper, K. (1959). *The logic of scientific discovery.* London: Hutchinson.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Ed.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519–559). Hillsdale, NJ: Erlbaum.

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

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power

have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.
Serlin, R. A., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.
Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70,* 434–438.
Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.
Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.
Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics, 33,* 1–67.
Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24,* 83–91.
Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.
Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6,* 100–116.
Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110.
Wainer, H., & Thissen, D. (1981). Graphical data analysis. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (pp. 191–241). Palo Alto, CA: Annual Reviews.
Wainer, H., & Thissen, D. (1993). Graphical data analysis. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical issues* (pp. 391–457). Hillsdale, NJ: Erlbaum.
Wilkinson, L. (1990). *SYGRAPH: The system for graphics.* Evanston, IL: SYSTAT.
Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association, 46,* 19–34.