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 H0 is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects H0 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 study is that at age three score and 10, originality is not which Bill Rozeboom (1960) wrote 33 years ago, "The theory that led to the test. Exploration 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.

What's wrong with NHST? Well, among many other things, it does not tell us what we want to know. And what we want to know is "Given these data, what is the probability that H0 is true?" But as most of us know, what it tells us is "Given that H0 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. Mehl, Stanley A. Mulaik, Robert Rosenthal, William W. Rozeboom, Elia Sinalko, Judith D. Singer, and Bruce Thompson. I also acknowledge the help I received from reviewers David Lykken, Matt McCue, 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.
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. The probability is .05.

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 an American, then he is not a member of Congress.

This person is a member of Congress. Therefore, he is not an American.

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

$H_0 = \text{The case is normal, so that}$

$H_1 = \text{The case is schizophrenia, and}$

$D = \text{The test result (the data) is positive for schizophrenia.}$

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...
null hypotheses. The confidence interval form. What I and my ilk decry is the form of the testing that has been used in astronomy and physics for centuries, what Meehl (1967) called the "weak" form in which theories are "confirmed" by replication. Ferguson, and Lindquist. Many examples of this authors making this error are Guilford, Nunnally, An-astasi. 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. Rot-enthal (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 H0 is rejected, then the theory is established: If A then 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 H0. Thus, when H0 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), H0 has little to commend it in the testing of psychological theories in its usual reject-H0-confirm-theory form. The ritual dichoto- mous 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-

<table>
<thead>
<tr>
<th>Result</th>
<th>Normal</th>
<th>Schiz</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative test (Normal)</td>
<td>949</td>
<td>1</td>
<td>950</td>
</tr>
<tr>
<td>Positive test (Schiz)</td>
<td>30</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>979</td>
<td>21</td>
<td>1,000</td>
</tr>
</tbody>
</table>

As the table shows, the conditional probability of a normal case for those testing as schizophrenics 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 H0 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, set up as the H0, 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 H0 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 H0, it is highly probable that replications of the research will also result in H0 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. Rotenthal (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 H0 is rejected, then the theory is established: If A then 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 H0. Thus, when H0 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), H0 has little to commend it in the testing of psychological theories in its usual reject-H0-confirm-theory form. The ritual dichoto-

mous 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-

December 1994 • American Psychologist
gree to which one . . . believes the hypothesis . . . being tested” (p. 420)

The Nil Hypothesis

Thus far, I have been considering \( H_0 \) 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 \( \phi \) 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 studied 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 measure with the admonition “Amount, as well as direction is vital” and went on to say the following:

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 “sancification” 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 example, Lee Wilkinson’s [1990] SYGRAPH, which is presently being updated).


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 intervals, which provide a more direct measure of the uncertainty in our estimates. Confidence intervals are calculated from the data and provide a range of values within which we can be confident the true effect lies. This approach is more informative than p-values, which only tell us whether the observed data are consistent with the null hypothesis.

In addition, we should also report confidence intervals for our estimates. This approach allows us to compare the precision of our estimates with those of other studies, and it provides a more direct measure of the uncertainty in our estimates. Confidence intervals are calculated from the data and provide a range of values within which we can be confident the true effect lies. This approach is more informative than p-values, which only tell us whether the observed data are consistent with the null hypothesis.

References


have an effect on the power of studies? *Psychological Bulletin.* 105, 309–316.


