Null Hypothesis Significance Testing

On the Survival of a Flawed Method

Joachim Krueger
Brown University

Null hypothesis significance testing (NHST) is the researcher's workhorse for making inductive inferences. This method has often been challenged, has occasionally been defended, and has persistently been used through most of the history of scientific psychology. This article reviews both the criticisms of NHST and the arguments brought to its defense. The review shows that the criticisms address the logical validity of inferences arising from NHST, whereas the defenses stress the pragmatic value of these inferences. The author suggests that both critics and apologists implicitly rely on Bayesian assumptions. When these assumptions are made explicit, the primary challenge for NHST—and any system of induction—can be confronted. The challenge is to find a solution to the question of replicability.

> Inductive inference is the only process known to us by which essentially new knowledge comes into the world. (Fisher, 1935/1960, p. 7)

> The supposition *that the future resembles the past* is not founded on arguments of any kind, but is derived entirely from habit, by which we are determined to expect for the future the same train of objects to which we have been accustomed. (Hume, 1739/1978, p. 184)

During my first semester in college, I participated in a student research project. We wanted to know whether people would be more willing to help a blind person than a drunk person in need. Using the wrong-number technique to collect data (Gaertner & Bickman, 1971) and a chi-square test to analyze them, we rejected the hypothesis that there was no difference in helping behavior. We learned from this experience that the analysis of experimental data leads to inferences about the probability of future events. When differences between conditions are improbable under the null hypothesis, researchers attribute these differences to stable underlying causes and thus expect to observe these differences again under similar circumstances. In Fisher's (1935/1960) words, "a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result" (p. 14).

Though plausible, the chain of inferences constituting null hypothesis significance testing (NHST) has often been criticized (see Morrison & Henkel, 1970, for an excellent collection of articles). Over the past decade, the debate over the validity of this method has become polarized with partisan arguments condemning its flaws (Cohen, 1994) or praising its virtues (Hagen, 1997). In search of common ground, I reviewed both attacks on NHST and the arguments brought to its defense, which ultimately led me to the same conclusion that Hume (1739/1978) drew more than 200 years ago: Inductive inferences cannot be logically justified, but they can be defended pragmatically.

Hume (1739/1978) observed that induction cannot be validated by methods other than induction itself: "There can be no demonstrative arguments to prove that those instances of which we have had no experience resemble those of which we have had experience" (p. 136). Induction from sample observations—no matter how numerous—cannot provide certain knowledge of population characteristics. Because induction worked well in the past, however, we hope it will work in the future. This itself is an inductive inference that can be justified only by further induction, and so on. Empirical research must either accept this leap of faith or break down. Because knowledge "must include reliable predictions" (Reichenbach, 1951, p. 89), we "act as if we have solved the problem of induction" (Dawes, 1997, p. 387).

Fisher (1935/1960) illustrated the properties of NHST with a test of Mrs. Bristol's claim that she could tell whether milk was added to tea or tea was added to milk. Following this example, I sometimes tell students that I can detect hidden objects. To test this claim, I ask a volunteer to hide a coin in one hand and to hold out both fists in front of him or her. Then I ask for the fists to be moved out to the sides, and I point to the one that I think holds the coin. Students may not believe that I am clairvoyant when I recover the coin, but they suspect that I have some relevant information. But why would they conclude anything after witnessing one successful demonstration? Assuming that Lady Luck grants success with a probability of .5, a single

Editor's note. J. Bruce Overmier served as action editor for this article.

Author's note. I am grateful to Melissa Acevedo, Robyn Dawes, Bill Heindel, Judith Schrier, Gretchen Therrien, and Jack Wright for their helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Joachim Krueger, Department of Psychology, Brown University, Box 1853, Providence, RI 02912. Electronic mail may be sent to joachim_krueger@brown.edu.

Copyright 2001 by the American Psychological Association, Inc. 0003-066X/01/$5.00
Vol. 56, No. 1, 16–26 DOI: 10.1037//0003-066X.56.1.16

Joachim Krueger

success is not "statistically significant." Students' apparent willingness to reject the luck hypothesis suggests that they perform an intuitive analogue of NHST with ã lax decision criterion.

Most scientists demand more evidence before attributing findings to something other than luck. Suppose I do the coin experiment eight times with seven successes. The probability of that happening, or anything more extreme (i.e., eight successes), is .035 if the null hypothesis is true. This result is obtained as the sum of the binomial probabilities for the number of successes (r) and any number more extreme (until $r = N$, the total number of trials). With p being the hypothesized probability of success on an individual trial, the formula is

$$\sum_{r}^{N} \binom{N}{r} p^{r}(1-p)^{N-r}.$$

NHST suggests that the chance (null) hypothesis can be rejected. This does not mean that clairvoyance has been proven. Less exotic explanations, such as trickery or sensitivity to nonverbal cues, remain. NHST simply suggests that the results need not be attributed to chance. It suggests that "there is not nothing" (Dawes, 1991, p. 252). Such an inference is a probabilistic proof by contradiction (modus tollens). If the null hypothesis is true, orderly data are improbable. If improbable data appèar, the null hypothesis is probably false. If the null hypothesis is false, then something else more substantive is probably going on (Chow, 1998).[1]

The key concern about this chain of inference is that deductive syllogisms are not valid when applied to induction. There are three specific criticisms. First, any point-specific hypothesis is false, and no data are needed to reject it. The goal of experimentation must therefore be something other than the rejection of null hypotheses. Second, even if one assumes that a hypothesis is true, data that are improbable under that hypothesis do not reveal how improbable the hypothesis is given the data. No contradiction, however, improbable, can disprove anything if the premises are uncertain. Third, significance levels say little about the chances of rejecting the null hypothesis in a replication study. NHST does not offer much help with predictions about future, yet-to-be-observed events. Defenders of NHST dispute each of these criticisms. I consider both sides of each argument and suggest possible resolutions.

The Null Hypothesis Is Always False: True or False?

Thesis: The Null Hypothesis Is Always False

In a probabilistic world, there is rarely "not nothing." Something is usually going on. Most human behavior is nonrandom, although little of it is relevant for the settling of theoretical issues. In a similar manner, any human trait is related to other traits by whatever small degree of association (Lykken, 1991). To show that there is not nothing does not make for rapid scientific progress (Meehl, 1990).

The argument that the null hypothesis is always false rests on the idea that hypotheses refer to populations rather than samples. Populations are mathematical abstractions assuming that the number of potential observations is infinite. An infinite number of observations implies an infinite number of possible states of the population, and each of these states may be a distinct hypothesis. With an infinite number of hypotheses, no individual hypothesis can be true with any calculable probability. "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)" (Cohen, 1990, p. 1308). If the probability of a point hypothesis is indeterminate, the empirical discovery that such a hypothesis is false is no discovery at all, and thus adds nothing to what is already known. A failure to detect the falsity of a hypothesis reflects only the imprecision of measurement or the limitations of sampling; it does not indicate that there is nothing to be detected in principle (Thompson, 1997).

If there is no expectation regarding the possible truth of the null hypothesis, its falsification by data is a redundant step. Falsification makes sense only when no exceptions are allowed. If one assumes, for example, that cows die when they are beheaded, a single surviving cow refutes this premise (Paulos, 1998). If, however, exceptions are allowed, no evidence can refute the hypothesis. Improbable data are just that: improbable. "With a large enough sam-

[1] These inferences characterize the weak use of significance tests, which is common in psychology. The strong use requires a substantive (non-nil) hypothesis to be subjected to potential falsification.

ple, any outrageous thing is likely to happen" (Diaconis & Mosteller, 1989, p. 859).[2]

Antithesis: Some Null Hypotheses Are True

Some defenders of NHST point out that the null hypothesis can be true in a finite population. Assuming error-free measurement, it is possible to show, for example, that exactly half of American men have fantasized about Raquel Welch. Because the number of American men is fixed at any given time, the null hypothesis can be true when this number is even. When the population does not have a fixed size, one would have to assume that it does. Assuming, for example, that a roulette wheel lasts 38 million spins, the null hypothesis is that each number (0, 00, and 1 through 36) comes up 1 million times.[3] A failure to reject the null hypothesis, given sample data, is then the correct decision. The question remains as to why the population should be limited to 38 million spins. Neither NHST nor any other formal mechanism solves this problem. There is no logical justification for predicating the presumed truth of the null hypothesis on a population of any particular size.

One pragmatic strategy is to estimate population size by relying on past experience. Tests of bias may be linked, for example, to the lifetimes of past roulette wheels. Although this strategy works well for casino operators, its logic remains circular. It justifies the validity of one inductive inference only by reference to another. In the coin-detection experiment, the null hypothesis is that I have no ability to locate the coin. If I decide on the total number of tests to be performed, I prejudge the decision about the truth or falsity of this ability. The more performances I anticipate, the smaller is the probability that exactly half will be successful. If, for example, I anticipate only 4 trials, the probability of two successes is .375; if I anticipate 10 trials, the probability of five successes is .246. When the number of anticipated trials approaches infinity, the probability of a match becomes infinitesimally small. Because the ability that I intend to test is an abstract idea, its existence cannot depend on the number of opportunities I have to exercise it. Once the population is allowed to be infinite, samples of any size can be drawn. Significance tests will eventually suggest that performance is either better or worse than chance.

"By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow the detection of . . . a quantitatively smaller departure from the null hypothesis" (Fisher, 1935/1960, pp. 21–22). Fisher's argument entails the impossibility of selecting a maximum number of observations without prejudging the status of the null hypothesis. It is impossible to claim that a sample is so large that its size is sufficiently similar to the population. Even the largest sample is infinitely smaller than the infinite population.

Intuitions about sample size contradict this claim. Some samples are so large that they seem to be representative of the population. Thus, the second argument against the assumed falsity of the null hypothesis points to notable failures to obtain significance (Oakes, 1975). Karl Pearson failed to reject the hypothesis that his coin was fair after 24,000 flips and 12,012 heads (Moore & McCabe, 1993). Instead of proving the null hypothesis, the small size of this effect—$p(\text{heads}) = .5005$—only predicts the persistence needed to make it significant. Significance eventually emerges because "whatever effect we are measuring, the best-supported hypothesis is always that the unknown true effect is equal to the observed effect" (Goodman, 1999b, p. 1007). If it were flipped four million times, Pearson's coin would probably be judged to be biased. Alas, practicing researchers are familiar with small effects that elude significance. The decision to leave them to nonsignificance is usually pragmatic, indicating that the estimated effect size does not justify the effort needed to attain significance.

The lack of significance does not establish the truth of the null hypothesis, however tempting this conclusion might be. Indeed, if there were one proven null hypothesis, the claim that all such hypotheses are false would itself be "demonstrably false" (Lewandowsky & Maybery, 1998, p. 210). There would be no telling how many more true null hypotheses there might be. Fisher (1935/1960) himself cautioned against attempts to prove the null hypothesis. Its falsity is, after all, an analytical matter, which cannot be verified by enumeration of rejected null hypotheses and which cannot be falsified by famous failures to reach significance.

The third argument defends NHST by allowing subjective beliefs to affect decisions about hypotheses. It says that some null hypotheses are true because we already know, or firmly believe, that they are true. Pearson assumed his coin to be fair, and the data did not strongly contradict his assumption. In a similar manner, skeptics adhere to null hypotheses because if they did not, they would have "to accept the fact that knocking on wood will prevent the occurrence of dreaded events, [or] that black cats crossing the road are better predictors of future mishaps than white cats [when] put to an experimental test with sufficiently large sample sizes" (Lewandowsky & Maybery, 1998, p. 210). This argument appeals to existing convictions that these things are not so. If tests with large

[2] Attempts to prove the logical validity of induction create only epistemic nightmares but no certainty. Hell, the incomparable Bertrand Russell (1955) imagined

> is a place full of all those happenings that are improbable but not impossible, [and that] . . . there is a peculiarly painful chamber inhabited solely by philosophers who have refuted Hume. These philosophers, though in hell, have not learned wisdom. They continue to be governed by their animal propensity toward induction. But every time that they have made an induction, the next instance falsifies it. This, however, happens only during the first hundred years of their damnation. After that, they learn to expect that an induction will be falsified, and therefore it is not falsified until another century of logical torment has falsified their expectation. Throughout eternity, surprise continues, but each time at a higher logical level. (p. 31)

[3] The inevitable rejection of the point-specific null hypothesis does not guarantee that the discerning player can enjoy betting on a favorable number. A number is favorable only if it comes up with a probability greater than 1/36 because 2 of the 38 numbers (0 and 00) yield no payoffs. In practice, therefore, this null hypothesis becomes a range hypothesis ($p < 1/36$; Ethier, 1982).

samples are found to be significant, the results would have to be Type I errors. In other words, the prior probability of the null hypothesis is so large that improbable data cannot easily threaten it.

Experiments with control conditions create an analogous situation. Random assignment to conditions without treatment ought not to produce differences in performance. Having tried to draw two samples from the same population, researchers assume that the null hypothesis is true. They have ruled out, as best they could, potential sources of differences between conditions. Like the belief in the fairness of a coin, however, the belief in perfectly random assignment is ultimately threatened by significant departures in very large samples. Reasoning pragmatically, most researchers therefore settle on the null hypothesis when it fails to be rejected by data from a finite sample. They act as if the null hypothesis is "true enough" for the purpose at hand.

From the practice of pragmatic acceptances of the null hypothesis, it is tempting to conclude that sometimes no increase in sample size—no matter how great—will lead to significance.

> Although it may appear that larger and larger Ns are chasing smaller and smaller differences, when the null is true, the variance of the test statistic, which is doing the chasing, is a function of the variance of the differences it is chasing. Thus, the "chaser" never gets any closer to the "chasee." (Hagen, 1997, p. 20)

The formula for the t statistic shows what this means. The index t is the difference between two means divided by the standard error of that difference. The standard error, in turn, is the standard deviation of the difference divided by the square root of the sample size. Thus,

$$t = \frac{D}{s/\sqrt{n}}, \text{ or } \frac{D\sqrt{n}}{s}.$$

Because D cannot be exactly 0 and because n has no ceiling, the test ratio will ultimately grow into significance. If the null hypothesis is postulated to be true, Hagen's argument is correct, but it begs the question of whether the null hypothesis is true. If the eventual emergence of significance is inevitable, why should any test be conducted at all? Although failures to reject the null hypothesis cannot prove anything, they may reveal the researchers' prior beliefs concerning the null hypothesis. Skeptics evaluating data regarding supernatural claims and experimenters evaluating data from control conditions accept the null hypothesis, in part, because they believe it to be true anyway.

The shortcoming of this objection (i.e., we know some null hypothesis to be true) is now clear. For mathematical reasons, which have nothing to do with the theoretical merit of the hypothesis, one will find that either a particular claim or its opposite has a kernel of truth. The color of cats (either black or not black) is related to the fate of those who encounter them. The association between these variables might well be ridiculously small, but a judgment about the ridiculousness of an effect size is not part of NHST. This judgment can be made only by a human appraising the size of the effect and the size of the sample necessary to coax this effect into significance. Most important, acceptance of nonzero associations between variables must be supported by plausible mechanisms (Goodman, 1999a). A small but significant correlation between the color of cats and the luck of their owners has little meaning unless something is known about the causes of this association. In a similar manner, the purpose of control conditions in experiments is to eliminate confounding variables. The identification of such variables, however, is a conceptual rather than a statistical matter.

Synthesis: Making the Subjective Element in Hypothesis Evaluation Explicit

Despite efforts to banish subjectivism from NHST, the practice of research shows how prior beliefs about the truth of hypotheses affect the subsequent evaluation of these hypotheses. This is hardly surprising because it is difficult to imagine how a hypothesis can be rejected without an implicit assessment of the improbability of the hypothesis given the evidence. Despite his opposition to inverse (i.e., Bayesian) probabilities, Fisher (1935/1960) understood that induction must enable us "to argue from . . . observations to hypotheses" (p. 3). Decisions about hypotheses refer to their posterior probabilities, $p(\mathrm{H}|\mathrm{D})$, and thus depend not only on the significance level, $\mathrm{p}(\mathrm{D}|\mathrm{H}_0)$, but also on the prior probabilities of the hypotheses, $p(\mathrm{H})$, and on the overall probability of the data, $p(\mathrm{D})$. Bayes's theorem states that

$$p(\mathrm{H}|\mathrm{D}) = p(\mathrm{H})\frac{p(\mathrm{D}|\mathrm{H})}{p(\mathrm{D})}.$$

The selection of hypotheses, their number, their location on the continuum of possible hypotheses, and their prior probabilities depend on the researchers' experience, their theoretical frame of mind, and the state of the field at the time of study. Consider three versions of the coin experiment in which observers entertain two different hypotheses regarding the probability of locating the coin on any individual trial. The null hypothesis, H_0, assumes performance at chance level ($p = .5$). Its complement, H_1, reflects a high skill level ($p = .9$).

The first scenario assumes that observers have no reason to favor either hypothesis before seeing the evidence. Professing ignorance, they assign the same prior probability to each. As I noted earlier, the probability of the data under the null hypothesis is .035. The probability of the data under the skill hypothesis is .81. The overall probability of the data is the sum of the two joint probabilities of hypothesis and data: $p(\mathrm{D}) = p(\mathrm{H}_0) \times p(\mathrm{D}|\mathrm{H}_0) + p(\mathrm{H}_1) \times p(\mathrm{D}|\mathrm{H}_1) = .42$. According to Bayes's theorem, the posterior probability of the null hypothesis is .04, and the posterior probability of the skill hypothesis is .96. The second scenario assumes that observers have some prior reason to believe that the coins will be found, perhaps because they have just heard a lecture on the use of nonverbal cues in person perception. If they assign a low prior probability to the null hypothesis ($p = .1$), its posterior probability is .005. The third scenario discourages expectations of success. When the coin searcher is blindfolded, for example, the null hypothesis appears to be rather probable ($p = .9$), and even seven successes out of eight trials leave a considerable posterior probability ($p = .28$).

The third scenario typifies "risky" research because the investigator doubts that the null hypothesis can be rejected. When such an experiment "works," the findings are impressive. A study is risky, for example, if the manipulation of its independent variable is only slight, or if the dependent variable is known to resist experimental influence (Prentice & Miller, 1992). Weak manipulations render the null hypothesis probable a priori, whereas strong manipulations make it improbable. Given identical evidence, Bayes's theorem suggests that the posterior probability of the null hypothesis remains higher after a weak manipulation than after a strong manipulation. The impressiveness of evidence is captured by the degree of belief revision, $p(H_0) - p(H_0|D)$, rather than by the strength of the posterior belief itself. Success in the coin experiment is more impressive with eyes closed than with eyes open.

Confusion About the Confusion of Probabilities

Thesis: Significance Says Little About the Rejectability of the Null Hypothesis

When only a limited number of hypotheses are being entertained, the first criticism of NHST is moot. The prior probability of the null hypothesis is assumed to be greater than zero, and it is therefore possible to estimate its posterior probability. In this situation, the critique of NHST turns to the validity of this estimate. Specifically, researchers are thought to ignore Bayes's theorem when deciding the status of the null hypothesis. Instead, they resort to fallible intuitions reminiscent of those found in everyday statistical reasoning. They conclude too readily that significant results imply the improbability of the null hypothesis.

Cohen (1994) offered a diagnostic example. Suppose that in tests of schizophrenia, the null hypothesis is that a person is normal, $p(H_0) = .98$. If the person is normal, the probability of a positive test result is .03, $p(D|H_0)$. Furthermore, the probability that schizophrenia is correctly identified is .95, $p(D|H_1)$. What the patient and the doctor need to know is the probability that a testee with a positive result is normal, that is, $p(H_0|D)$. Bayes's theorem reveals this probability to be .61. People who ponder problems like this tend to underestimate this probability. They consider the null hypothesis to be unlikely when the data are unlikely under that hypothesis. In Cohen's example, inferences about the testee's health status depend too much on the false-positive rate of the test (here, .03) and too little on the probability of health regardless of the test (here, .98).

Falk and Greenbaum (1995) presented many examples of authors, reviewers, editors, and textbook writers wrongly believing that the null hypothesis is rendered improbable (i.e., rejectable) by evidence that is improbable under that hypothesis (see also Bakan, 1966; Carver, 1978; Gigerenzer, 1993; Oakes, 1986). Hays and Winkler (1971), for example, wrote that "a p-value of .01 indicates that H_0 is unlikely to be true" (p. 422). Why do many researchers rush to reject the null hypothesis? The most obvious reason is that Fisher's (1935/1960) method seduces practitioners to make decisions about the null hypothesis on the basis of incomplete information. According to Fisher, "every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (p. 16). If $p(D|H_0)$ is all the method provides, how are researchers supposed to reach a decision concerning the falsity of the null hypothesis if not by using $p(D|H_0)$? If researchers suspended judgment, citing the incompleteness of the information, they could not justify why they ran the experiment in the first place.

Numerous heuristics and biases have been shown to affect probabilistic reasoning in everyday contexts. These reasoning shortcuts may also guide the researchers' inference processes. The heuristic of anchoring and insufficient adjustment suggests that probability estimates are biased by whatever number is offered as potentially relevant, even if that number is exposed as arbitrary (Tversky & Kahneman, 1974). When a low significance level is the only available anchor, the estimate for $p(H_0|D)$ is easily distorted. Heavy reliance on significance levels is also consistent with the representativeness heuristic. Because the two inverse conditional probabilities appear to be conceptually similar, people assume that $p(H_0|D) = p(D|H_0)$. But as Dawes (1988) noted, "Associations are symmetric; the world in general is not" (p. 71).

Gigerenzer (1993) offered a tongue-in-cheek Freudian metaphor. Although the frequentist superego forbids it, the Bayesian id wants to reject the null hypothesis on the basis of improbable evidence. The pragmatic Fisherian ego allows the id to prevail because otherwise nothing is accomplished (i.e., published). This neurotic arrangement is supported by social factors such as rigid training in the rituals of NHST and the stated policies of journal editors.

Antithesis: Though Illogical, NHST Works in the Long Run

The charge that null hypotheses are tossed out too easily need not mean that NHST must be abandoned. Rejecting null hypotheses may be better than doing nothing. This view echoes Hume's (1739/1978) conclusion that induction is useful if it is properly understood as a matter of custom and habit rather than logic. Induction may not work, but it will if anything does (Reichenbach, 1951). I consider two specific defenses for the use of significance levels in decisions about hypotheses.

The first argument is that a Bayesian critique of NHST lacks an objective foundation. Most prior probabilities of hypotheses are subjective; unlike significance levels, they cannot be expressed as long-range frequencies. Because posterior probabilities are derived, in part, from these prior probabilities, they have no objective status either. When making decisions regarding the presumed truth or falsity of the null hypothesis, researchers only act as if they are expressing a posterior probability. When forced, perhaps against their better instincts, to estimate the posterior probability of the null hypothesis, researchers may just assume that $p(H_0|D) = p(D|H_0)$. This assumption permits the overall probability of the data, $p(D)$, and the prior probability of the null hypothesis, $p(H_0)$, to assume any value as long as the two are the same. The claim that $p(H_0|D)$ is always the same as $p(D|H_0)$ is thus empirically sterile, and no evidence can refute it.

The second argument in favor of significance levels is that they minimize decision errors in the long run. Across studies, significance levels are correlated with the posterior probabilities of the null hypothesis for any given set of prior probabilities (Dixon, 1998; Hagen, 1997). An experiment with a $p(D|H_0)$ of .01 provides stronger evidence against the null hypothesis than an experiment with a probability of .05. In coin-detection experiments with eight trials and $p(H_0) = .5$, for example, six and seven successes yield $p(D|H_0) = .145$ and .035, respectively. With the skill hypothesis still assuming that p (success) = .9 on each trial, the posterior probability of the null hypothesis is smaller after seven (.04) than after six (.17) successes.

Synthesis: Another Look at the Association Between Inverse Conditional Probabilities

When there is no expectation of what the size of the effect might be, the effect obtained in any initial study of a phenomenon is the best estimate of H_1. Assuming this alternative to the null hypothesis is true, the probability of the obtained data, or data more extreme, is .5 (or a probability very close to it; Hedges, 1981). In practice, however, there is no assurance that the alternative hypothesis is true. Recall that researchers often distinguish between risky experiments, in which the null hypothesis is probable, and safe experiments, in which it is improbable. It is therefore useful to consider a range of prior probabilities to fully examine the relationship between significance levels and the posterior probabilities of the null hypothesis. When posterior probabilities are plotted against varying significance levels, they teach two lessons (see Figure 1). First, the prior probability of the null hypothesis has to be less than .35 for the posterior probability of the null hypothesis to be as low as the conventional significance level of .05. Second, $p(D|H_0)$ and $p(H_0|D)$ are correlated ($r = .38$), with their association best described by a logarithmic function, .0852 ln (x) + .4107.

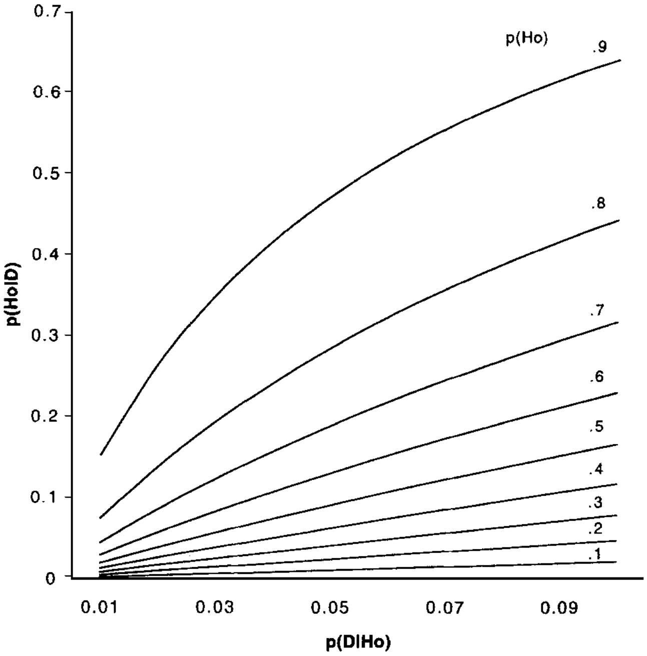
In a way, the critics and the defenders of NHST are talking at cross-purposes. Whereas the critics emphasize the inequality of inverse conditional probabilities, unless $p(H) = p(D)$, defenders point out that Bayes's theorem guarantees the two probabilities to be correlated. Again, a solution to this controversy lies in making prior beliefs about hypotheses explicit.

The Heart of Induction: Replicability

Thesis: NHST Says Little About the Replicability of Results

If practicing scientists agree on anything, it is that evidence must be replicable. If the data collected in the past say nothing about data to be gathered in the future, empirical research is merely historical. A null hypothesis that was rejected once needs to be rejected again because "no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon" (Fisher, 1935/1960, p. 14). Replicability is the probability that the null hypothesis will be rejected in a follow-up study given that it was rejected in an initial study. If seven out of eight predictions were correct in the coin experiment, one may ask how likely it is that this feat will be repeated.

Figure 1
Posterior Probabilities of the Null Hypothesis (H_0) as a Function of Significance Levels



Note. D = data.

The most serious critique of NHST is that "a small p-value [is] *unrelated to the question of reliability*" (i.e., replicability; Falk & Greenbaum, 1995, p. 90). Gigerenzer (1993) remained agnostic about the question of replicability, saying that "the answer is unknown" (p. 329). But what is the purpose of experiments if the replicability of their results cannot be estimated? "What is knowledge if it does not include the future" (Reichenbach, 1951, p. 89)? Falk and Greenbaum suggested that the question of replicability can be settled only by replication itself:

> Instead of measuring the quality of research by the level of significance, it would be better judged by its consistency of results in repeated experiments [and] if a researcher does obtain the same result . . . more than once, it strengthens the conclusion that the results are not due to chance. (p. 92)

Successful replications push the null hypothesis further toward improbability. Experiments with a lot of statistical power do the same thing, however. Both replications across studies and significance levels in individual studies are events of the past, and both predict the probability of future replications. If it were impossible to estimate replicability from single studies, it would also be impossible to do so from many studies. The distinction between the significance level of a single study and the record of replications in the past is specious. The force of the evidence differs only quantitatively, but not in essence.

A less serious critique is that the chances of replication can be estimated but that practitioners of NHST routinely overestimate replicability (Tversky & Kahneman, 1971). Many research psychologists believe, for example, that replicability is the inverse of the significance level (Oakes, 1986). This "cognitive trap" (Falk & Greenbaum, 1995, p. 90) follows from the confusion of inverse conditional probabilities. If researchers conclude from a significance level of .05 that the null hypothesis has a probability of .95 of being false, they might also conclude that the same hypothesis has a probability of .95 of being rejected the next time around. This fallacy is well documented. It is an educational rather than a methodological problem.

Antithesis: NHST Foretells Replicability

Greenwald, Gonzalez, Harris, and Guthrie (1996) showed that "a p value resulting from NHST is monotonically related to an estimate of a non-null finding's replicability" (p. 179). This relationship holds when three assumptions are met. The first assumption is that only two hypotheses, H_0 and H_1, are in contention. The second assumption is that H_1 is identified post hoc with the effect size observed in the initial study (see also Goodman, 1999b; Hagen, 1997; Krueger, 1998). The third assumption is that replicability is "the power of an exact replication study" (Greenwald et al., 1996, p. 179). In other words, replicability is understood as the probability of the data (or data more extreme) under the alternative hypothesis, that is, $p(D|H_1)$. If, for example, the significance level of an initial study is .05, the probability of finding significance in an exact replication study is .5. If, however, the initial significance level is more extreme (p = .01), a successful replication ($p < .05$) is more probable. In a z distribution, power can be estimated by subtracting the z score for the minimum desired significance level (e.g., 1.96 for p = .05, two-tailed) from the z score for the obtained level (e.g., 2.58 for p = .01) and by finding the cumulative probability of the difference (p = .73). If, for example, seven out of eight trials are successful in the initial coin experiment (p = .035), and if an ability level of 7/8 is considered the most likely hypothesis, the probability of rejecting the null hypothesis by getting at least seven successes in Experiment 2 is .74.

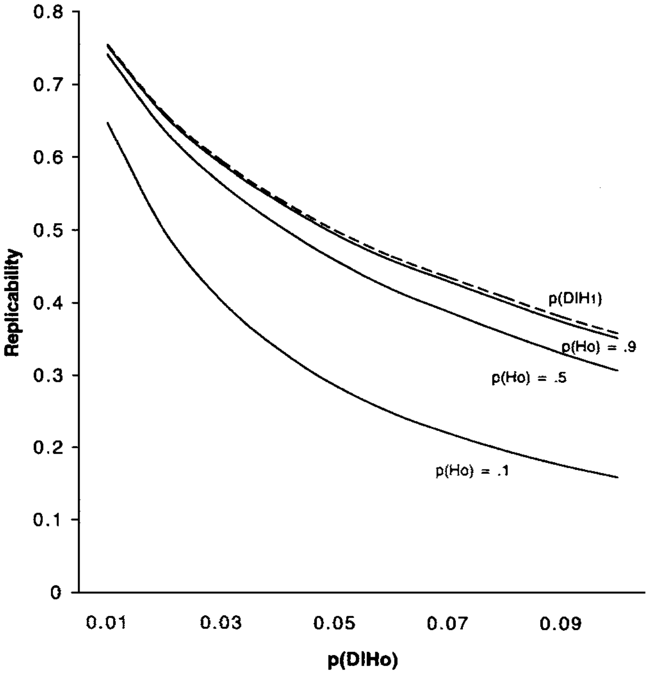
Synthesis: Bayesian Assumptions Are Vital for the Estimation of Replicability

The method proposed by Greenwald et al. (1996) estimates the replicability of a finding as the probability of rejecting the null hypothesis assuming that the alternative hypothesis is true. But recall that the truth of the alternative hypothesis is only a good guess. The null hypothesis has been rejected but not disproven. Each hypothesis has a posterior probability, which depends in part on its prior probability. To estimate replicability, it is necessary to estimate both the probability of rejecting the null hypothesis if the alternative hypothesis is true (power) and the probability of rejecting the null hypothesis if the alternative hypothesis is false (Type I error). The sum of these two probabilities is the probability of replication.

In the first scenario of the coin experiment (uniform prior probabilities), the prior probability of seven or eight successes is .42. After seven successes, the posterior probability of each hypothesis is multiplied by the probability of this result given that hypothesis. The sum of the products is the probability that a new experiment will yield at least seven successes, $p(D)$ = .78. In the eyes-open scenario, $p(D)$ increases from .74 to .81, and in the eyes-closed condition (Scenario 3), it increases from .11 to .60. If, in contrast, the power of the next experiment is accepted as the probability of replication, expectations are high regardless of the prior probabilities. After seven successes, the probability of seeing at least seven successes, assuming that the skill hypothesis is true, is .81.

For a more comprehensive view of the relationship between significance levels and the replicability of the data (with $p < .05$), consider a situation with two hypotheses and data yielding z scores. The prior probability of the null hypothesis is .1, .5, or .9, with each prior probability of the alternative hypothesis being the complement. Figure 2

Figure 2
Replicability as a Function of Significance Levels



Note. D = data; H_1 = alternative hypothesis; H_0 = null hypothesis.

shows coefficients of power and replicability plotted against nine significance levels. The graph illustrates four points. First, power coefficients decrease as significance levels increase ($r = -.97$). Second, replicability coefficients also decrease ($r = -.97$ for each $p[H_0]$). Third, replicability coefficients are lower than their respective power coefficients regardless of prior probabilities and significance levels. This difference appears because a rejected null hypothesis may have been false—as assumed by power analysis—or true. Fourth, the discrepancies between the replicability coefficients for different prior probabilities diminish as significance levels decrease. Evidence eventually overwhelms a priori differences in opinion because the prior probabilities enter the inference chain only at the beginning of the research process (Lindley, 1993). Evidence can accumulate indefinitely and continue to reduce significance levels.

These examples show that significance levels are useful cues for the replicability of empirical phenomena. If this is what practitioners learn from their work, they reason well. An editor of the *Journal of Experimental Psychology* once suggested that studies rejecting the null hypothesis at .01 should be preferred for publication because of the high reliability (i.e., replicability) of their findings (Melton, 1962). Melton's editorial has often been cited as an example of poor statistical reasoning. In contrast, the present analysis suggests that although Melton's policy lacked a logical foundation, it was reasonable in a pragmatic sense.

Single Studies, Replications, and Meta-Analyses

I now return to Falk and Greenbaum's (1995) suggestion that the demonstrability of a phenomenon hinges on the question of whether initial findings are actually replicated. If so, confidence in the falsity of the null hypothesis can be boosted only by rejecting the null hypothesis in a second study, not by obtaining a lower significance level in the first study. In contrast, the foregoing analysis showed that significance levels forecast replicability, just as separate successful replication studies do.

Exact replications do not provide information that cannot be gained from single studies with large samples; they just more rapidly decrease p values. Consider a replication of the coin experiment. Under the null hypothesis, the probability that both experiments yield at least 7 successes is $.035^2 = .0012$. But suppose the researcher runs a single experiment with 16 trials rather than two experiments with 8 trials each. The probability that such an experiment yields at least 14 successes is .002 under the null hypothesis. This significance level is not as low as the combined level from two studies because the 2 failures can occur anywhere in the sequence of 16 trials. There might be 1 failure in the first half of the experiment and 1 in the second half, or both might be in the same half. In the case of two successful 8-trial studies, however, neither one can contain 2 failures.

Consider the consequences of randomly breaking up the 16-trial study into two separate experiments. Such a post hoc split is legitimate when both samples are drawn from the same population. What is the probability that this method yields two rejections of the null hypothesis? A successful replication requires that the two failures are located in different post hoc experiments. Once the location of one failure is known, the probability that the other is in the same experiment is .467 because there are seven other possible locations in the same experiment but eight possible locations in the other. Thus, it is slightly more likely that a post hoc split results in one rejection of the null hypothesis (with $p = .004$) and one nonrejection (with $p = .145$). Only in the unlikely case that there is one or no failure among the 16 trials ($p = .0003$) does any post hoc split yield two rejections of the null hypothesis. Successful replications have no special advantage over single studies. Unsuccessful replications, however, leave problems of interpretation for researchers accustomed to tallying up the rejection–nonrejection record of a research program. Meta-analyses avoid this nose-count strategy by combining significance levels across studies, with the familiar result that many replications relegate the data (and thereby the null hypothesis) to the remote reaches of improbability.[4]

Replication studies are most valuable if they are conceptual rather than exact (Lykken, 1968). Conceptual replications are risky because they involve variations in method and design while preserving the substantive hypothesis. The hypothesis that certain nonverbal signs reveal the location of the coin, for example, can be tested further by recruiting a different target person or by training a different observer.[5] Researchers gain more confidence in the robustness of a phenomenon after a successful conceptual replication than after an exact replication. After a conceptual replication, it is difficult to quantify, however, just how improbable the combined evidence has become under the null hypothesis, and so the gain in confidence is qualitative rather than quantitative (Dawes, 1997).

Induction Without Compunction

This review leads to three general conclusions: First, a logical analysis reveals irreparable limitations of NHST. No deductive syllogism can establish the validity of this method, nor can this method validate itself. Second, NHST rewards the pragmatic scientist. Much has been learned from this method in the past and presumably more can be learned in the future. Third, the rift between the logical incompleteness of NHST and its pragmatic value may be understood, in part, on psychological grounds (see also Harlow, Mulaik, & Steiger, 1997).

[4] Even nonsignificant replications lower the combined p value as long as results are in the same direction. The original significant p value is multiplied with another probability that is smaller than 1. Eventually, even series of nonsignificant results become meta-analytically significant.

[5] To make a conceptually broader case for the claim that trained observers can accurately interpret subtle nonverbal signs without knowledge of those who emit these signs, the type of sign should be varied as well. As noted conjuror and cognitive psychologist Ray Hyman (personal communication, June 1988) suggested, the observer can close his eyes and ask the participant to briefly raise the hand holding the coin. When the observer opens his eyes after the hand has been lowered again, he can locate the coin by noting which hand is paler.

The Limits of Knowledge

Induction is the admission that we cannot know everything. We generalize not only to events that we hope to observe in the future but also to observations that we expect to remain unrealized. Fisher (1935/1960) knew that induction must involve leaps from a known past to an uncertain future. At times, however, he denied uncertainty, suggesting that the goal of his method was "to supply the machinery for unambiguous interpretation" (Fisher, 1935/1960, p. vii). Much of the confusion created by NHST lies in the fact that the inferential aspect of this method is distinct from its computational aspect (Goodman, 1999a). The computational machinery guarantees that researchers following the same analytical recipe will extract the same information—namely, $p(D|H_0)$—from the same data. When Fisher referred to the "logic" of experimentation, he may have had this objective aspect in mind. In contrast, inferences remain subjective and thus ambiguous. Although researchers may agree to reject the null hypothesis when the probability is less than .05, their theoretical conclusions may diverge because they (a) implicitly consider different prior probabilities, (b) are prone to biases in estimating inverse probabilities or probabilities of replication, or (c) differ in their skill or persistence to gather more data. As one eminent statistician cautioned, "*There is no God-given rule about when and how to make up your mind in general*" (Hays, 1973, p. 353).

The ambiguities of the inference stage can be reduced when researchers' Bayesian assumptions are explicit. As a final illustration of this central point, consider researchers' affective response to their findings. Fisher (1935/1960) thought that the value of NHST lies in its capacity to express "the nature and the degree of the uncertainty" (p. 4). More extreme significance levels increase certainty in the falsity of the null hypothesis. Ironically, however, the emotion of surprise is said to increase at the same time. "We can regard the statistical significance or nonsignificance of a result as a measure of the 'surprisal value' of the result" (Hays, 1973, p. 384). Or, more precisely, "a *p* value's measure of surprise is simply captured by the number of consecutive zeros to the right of its decimal point" (Greenwald et al., 1996, p. 178).

The contradiction between increasing certainty and surprise disappears when one assumes that they occur sequentially. Giving the null hypothesis some credence a priori, researchers are surprised to see it rendered improbable, and they then gain certainty about its falsity. More important, the prior probability of the null hypothesis moderates these affective reactions. Safe experiments elicit greater certainty and less surprise than risky experiments do. In short, the researchers' emotional reactions on seeing their data reveal their implicit assumptions.

Progress Despite Limits

In daily research activity, NHST has proven useful. Researchers make decisions concerning the validity of hypotheses, and although their decisions sometimes disagree, they are not random or arbitrary. Empirical research clearly requires some mechanism for induction. NHST, especially when bolstered by Bayesian assumptions, fares quite well relative to its alternatives. Effect sizes alone do not weed out findings with large sampling errors (Frick, 1996). Significance levels help do this; they indicate how hard a researcher has worked to reduce error to at least make a judgment about the direction of the effect.[6] Significance levels also forecast replicability better than effect sizes do (Greenwald et al., 1996). This benefit is not surprising because *p* depends on both the effect size and the sample size, whereas the effect size—by definition—is only that: the effect size. When a small effect is significant, NHST does not reveal whether this effect is worth reporting. Judgments of relevance require the consideration of subjective values (i.e., costs and benefits). Ideally, these values are consulted beforehand so that minimum effect sizes can be determined. Unfortunately, researchers are often unwilling or unable to estimate desired or required minimum effect sizes. Finally, confidence intervals are attractive because they contain more information than *p* values do, but they steer researchers away from drawing categorical conclusions about hypotheses (which is what they are trained to aim for; Feinstein, 1998).

Given the pragmatic benefits of NHST and the lack of a superior alternative, an all-out ban of this method (Hunter, 1997) seems unnecessary. The American Psychological Association suggests that inferential tests be supplemented by other statistical indices (e.g., effect size measures; Wilkinson & the Task Force on Statistical Inference, 1999). In short, the pragmatic benefits of induction explain the continued popularity of NHST. When predicting one's own future behavior, it is better to assume that it will resemble one's past behavior than to assume nothing (but see Dawes, 1988). Thus, I continue to use NHST when analyzing data. Kaplan and I recently reported that under certain conditions, a threat to a person's self-concept can enhance responsiveness to social influence (Kaplan & Krueger, 1999). That research was rewarding because it reminded me of my first contact with psychological experimentation (recall the experiment with the wrong-number technique), and once again, a significant chi-square statistic suggested that something interesting (i.e., not nothing) was going on.

Researchers will probably continue to use NHST to draw inferences beyond the data given. Even the staunchest critics of NHST publish articles with conclusions stemming from rejected null hypotheses (Dawes, 1997; Greenwald et al., 1996). It is not entirely clear, however, why they do this. One explanation is that the critics recognize the pragmatic benefits of NHST, but if they do, one wonders why they bother to reveal its logical pitfalls. Another possibility is pluralistic ignorance. Perhaps critics underestimate how

[6] In areas considered to be soft in psychology (abnormal, personality, social), effect sizes are negatively correlated with sample sizes but are positively correlated with power coefficients (Sedlmeier & Gigerenzer, 1989). Many tests of subtle phenomena involve just enough observations to yield statistical significance. This suggests that at least some researchers sequentially test null hypotheses and stop collecting data once significance is obtained.

effective their attacks have been. Thinking that most other researchers have remained committed to NHST, they, the critics, analyze data and report results in a way they think will yield the greatest dissemination. In a related manner, critics may remain wary of editors who enforce traditional statistical rituals.

To Every Thing There Is a Season

If NHST is here to stay, how much longer will it last? This is a delicious question because it requires a forecast for the lifetime of an inductive method, which can be made only on the basis of some other inductive method. The Copernican principle states that most events are commonplace rather than special (Gott, 1993). When applied to a point in time (i.e., now) within a finite period, this point is probably not near the beginning or the end of this period. Therefore, the distance to the beginning (looking back) predicts the distance to the end (looking ahead). Things that only recently came into being (e.g., the Berlin Wall, as Gott [1993] predicted in 1969) tend to come to an end before things that have been around for a while (e.g., the Great Wall of China).

If Fisher's first edition of *The Design of Experiments* in 1935 is taken as the onset of the NHST era, a 50% confidence level is bounded by a minimum of another 22 years (25% of its lifetime is left) and a maximum of 195 years (25% of its time has passed). Although this interval is large, it provides a more reasonable prediction than hasty intuitions about the end being near. Copernican predictions provide good guesses for unique events, such as the lifetime of the Berlin Wall, the Soviet Empire, or the survival of *homo sapiens.* For events that belong to categories with well-understood lifetimes, however, it is better to rely on the base rates of survival. It would be foolish, for example, to expect an 80-year-old to live longer than an 8-year-old or to expect an old car to outlast a new one.

Induction will probably be around for a while, and Hume (1739/1978) will be cited for more years than Fisher (1935/1960). To Hume, induction was a habit of the mind providing a bridge from observation to learning. Incidentally, the logical insufficiency of this bridge applies to Hume's own argument. He could not be certain that people would continue to learn by induction, but now as much as then, this expectation is a good one.

REFERENCES

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

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

Chow, S. L. (1998). Statistical significance: Rationale, validity and utility. *Behavioral and Brain Sciences, 21,* 169–240.

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

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

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

Dawes, R. M. (1991). Probabilistic versus causal thinking. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Vol. 1. Matters of public interest: Essays in honor of Paul Everett Meehl* (pp. 235–264). Minneapolis: University of Minnesota Press.

Dawes, R. M. (1997). Qualitative consistency masquerading as quantitative fit. In M. L. Dalla Chiara, D. Kees, D. Mundici, & J. van Bentheim (Eds.), *Structures and norms in science* (pp. 387–394). Dordrecht, the Netherlands: Kluwer Academic.

Diaconis, P., & Mosteller, F. (1989). Methods of studying coincidences. *Journal of the American Statistical Association: Applications & Case Studies, 84,* 853–861.

Dixon, P. (1998). Why scientists value *p* values. *Psychonomic Bulletin & Review, 5,* 390–396.

Ethier, S. N. (1982). Testing for favorable numbers on a roulette wheel. *Journal of the American Statistical Association: Theory and Methods, 77,* 660–665.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard. *Theory & Psychology, 5,* 75–98.

Feinstein, A. R. (1998). *P*-values and confidence intervals: Two sides of the same unsatisfactory coin. *Journal of Clinical Epidemiology, 51,* 355–360.

Fisher, R. A. (1960). *The design of experiments* (8th ed.). Edinburgh, Scotland: Oliver & Boyd. (Original work published 1935)

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1,* 379–390.

Gaertner, S., & Bickman, L. (1971). Effects of race on the elicitation of helping behavior: The wrong number technique. *Journal of Personality and Social Psychology, 20,* 218–222.

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

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine, 130,* 995–1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine, 130,* 1005–1013.

Gott, J. R. (1993). Implications of the Copernican principle for our future prospects. *Nature, 363,* 315–319.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175–183.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

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

Hays, W. L., & Winkler, R. L. (1971). *Statistics: Probability, inference, and decision.* New York: Holt, Rinehart & Winston.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.

Hume, D. (1978). *A treatise of human nature.* Glasgow, Scotland: William Collins. (Original work published 1739)

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 3–7.

Kaplan, A., & Krueger, J. (1999). Compliance after threat: Self-affirmation or self-presentation? *Current Research in Social Psychology, 4,* 178–197. Retrieved December 14, 2000, from the World Wide Web: http://www.uiowa.edu/~grpproc/crisp/crisp.4.7.htm

Krueger, J. (1998). The bet on bias: A foregone conclusion? *Psycoloquy, 9*(46). Retrieved December 14, 2000, from the World Wide Web: http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?9.46

Lewandowsky, S., & Maybery, M. (1998). The critics rebutted: A Pyrrhic victory. *Behavioral and Brain Sciences, 21,* 210–211.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics, 15,* 22–25.

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

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Vol. 1. Matters of public interest: Essays in honor of Paul Everett Meehl* (pp. 3–39). Minneapolis: University of Minnesota Press.

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

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64,* 553–557.

Moore, D. S., & McCabe, G. P. (1993). *Introduction to the practice of statistics.* New York: Freeman.

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

Oakes, W. F. (1975). On the alleged falsity of the null hypothesis. *Psychological Record, 25,* 265–272.

Oakes, W. F. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* Chichester, England: Wiley.

Paulos, J. A. (1998). *Once upon a number.* New York: Basic Books.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160–164.

Reichenbach, H. (1951). *The rise of scientific philosophy.* Berkeley: University of California Press.

Russell, B. (1955). *Nightmares of eminent persons.* New York: Simon & Schuster.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Thompson, B. (1997). In praise of brilliance: Where that praise really belongs. *American Psychologist, 52,* 799–800.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110.

Tversky, A., & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist, 54,* 594–604.