Null-hypothesis significance testing (NHST) has been severely criticized because it involves deciding between two alternatives, one of which (the null hypothesis) is known a priori to be false for any but contrived data. However, as Tukey (1991) pointed out, NHST is not really a matter of deciding whether or not \( H_0 \) is true. Rather, we test \( H_0 \) because if we cannot rule out the possibility that our population effect is exactly zero, we also can not rule out both small positive and small negative values of that effect, and therefore cannot be sufficiently confident that we know its sign (direction). Indeed, as applied by most researchers and journal editors, NHST provides a very useful form of social control over researchers’ understandable tendency to squander analytic effort “explaining” effects whose sign in a given sample may not match the sign of the corresponding population effect. This contribution must be retained in any replacement system.

THREE-VALUED (SENSIBLE) HYPOTHESIS-TESTING LOGIC

Kaiser’s (1960) presentation of NHST as involving three (not just two) alternatives makes the logical basis of this contribution clear. Consider Figure 1’s presentation of three-valued logic for the two-sample test (two-tailed, split-tailed, and one-tailed versions thereof).

Beginning with the top panel, we can summarize the research behavior implied by a three-alternative two-tailed test as follows. If the data provide strong evidence that \( \mu_1 > \mu_2 \), that is what we conclude. If the data provide strong evidence that \( \mu_1 < \mu_2 \), that is what we conclude. And if the evidence for a particular direction of effect is not compelling, we conclude that we have insufficient evidence to conclude whether \( \mu_1 > \mu_2 \) or vice versa.

To provide more power for detecting a difference in the predicted direction than for detecting a difference opposite to prediction, we can adopt Braver’s (1975) split-tailed test (middle panel). For example, we might set \( \alpha_\text{up} \) (the proportion of the area in the predicted tail of the null distribution) to .04 and \( \alpha_\text{down} \) (the proportion in the nonpredicted tail) to .01.

Setting \( \alpha_\text{up} \) to zero yields a one-tailed test. As the bottom panel in Figure 1 makes clear, a one-tailed test (a) is simply an infinitely biased split-tailed test that can never lead us to conclude that we are wrong about the direction of the population effect and (b) is therefore never appropriate in a research setting.

Finally, we never conclude that \( \mu_1 \) does equal \( \mu_2 \) exactly because the difference between the sample means (our best available estimate of the population difference) is almost never exactly zero. There are no true null hypotheses except by construction.

TWO-VALUED (TRADITIONAL, BUT SILLY) LOGIC

NHST as just presented is consistent with sound research practice. However, it can be (and is) employed in this way only by ignoring the way in which its logical structure is presented in almost all of our textbooks—namely, as a choice between two mutually exclusive hypotheses. Consider Figure 2’s generic representation of the two-valued logic almost universally offered (but see Harris, 1994) as the basis for NHST.

In this logic, both the right- and the left-hand rejection regions for the two-tailed and split-tailed tests are labeled “\( \mu_1 \neq \mu_2 \)” As can readily be seen by scanning the rejection-region labels in this figure, a researcher who adheres to the two-valued logic of the traditional two-tailed test can never come to any conclusion about the direction of the population effect, and a researcher who instead employs a one-tailed test can never come to the conclusion that the predicted sign of the effect is wrong (i.e., that, under the conditions of this particular study or meta-analysis, the effect goes in the direction opposite to prediction). The researcher who takes traditional NHST logic seriously is thus faced with an unpalatable choice between (a) being unable to come to any conclusion about the sign of the effect or (b) violating the most basic tenet of the scientific method (empirical data as the arbiter of our conclusions) by declaring this directional hypothesis impervious to disconfirmation.

The contradictions between sound research practice and two-valued hypothesis-testing logic lead some of our colleagues to such absurdities as stating whether or not results are statistically significant, but not in what direction; stating hypotheses in the introduction only in null form; and advocating or carrying out one-tailed tests of research hypotheses. These contradictions also lend credibility to our students’ impression that statistical procedures are merely an arbitrary series of barriers to publishing one’s research.

Adopting three-valued hypothesis-testing logic would require no changes in the conduct of scientifically appropriate research or the reporting thereof, but only changes in the way in which we describe the underlying logic of NHST in textbooks, to our colleagues and consultees, and to ourselves. I believe, however, that misinterpretations of NHST would be much less likely under three-valued logic than under two-valued logic.
CONFIDENCE INTERVALS VERSUS (?) SIGNIFICANCE TESTS

Confidence intervals provide additional information about the range of possible values of \( \mu_1 - \mu_2 \) and thus provide humility when our difference is statistically significant and help us distinguish between low power and small population effect size as the reason for a nonsignificant \( t \). However, there are some uses for which significance tests are equivalent to or superior to confidence intervals.

Equivalence of Error Rates for NHST and Confidence Intervals

Schmidt (1996) pointed out that confidence intervals (the preferred alternative to NHST in single studies) can also be used to make the same decision between \( H_0 \) and \( H_1 \). This follows from the fact that the \( t \) test of \( M_1 \) versus \( M_2 \) at the .05 level is statistically significant if and only if the .95 confidence interval around \( \mu_1 - \mu_2 \) does not include zero.

Because use of a significant test and use of a confidence interval lead to rejection or nonrejection of \( H_0 \) in exactly the same cases, these two methods must have the same probability of rejecting \( H_0 \) (Type I error if \( H_0 \) is true, power otherwise), of failing to reject \( H_0 \) (Type II error), and of rejecting \( H_0 \) in the wrong direction (Type III error) when used for this sign-of-population-effect purpose. If we conduct a study with only 40% power (a level commonly used in NHST-ban proponents' arguments), then the confidence-interval-based procedure will have the same 60% probability of failing to provide a conclusion as to the sign of our effect that NHST does. The claim that the overall error rate for confidence intervals is only 5% comes from comparing two very different kinds of error. A large percentage of the "correct" 95% of these statements will be vacuously correct, that is, will include both positive and negative values of \( \mu_1 - \mu_2 \) and will thus fail to provide a decision as to the sign of the population effect, making even the most basic comparison between competing theories impossible.

If, then, we wish to retain in a confidence-interval-based alterna-

---

**Fig. 1.** Rejection and nonrejection regions under three-valued (sensible) hypothesis-testing logic (Kaiser, 1960). For the split-tailed and one-tailed tests, it is assumed that the research hypothesis is that \( \mu_1 > \mu_2 \).

---

**Fig. 2.** Rejection and nonrejection regions under two-valued (traditional, but silly) hypothesis-testing logic (generic). For the split-tailed and one-tailed tests, it is assumed that the research hypothesis is that \( \mu_1 > \mu_2 \).
Significance Tests Have Their Place

tive to significance testing the kind of social control over (over)interpretation of our data that NHST has provided over the past several decades, we will find that the error rates of the alternative system in this respect are identical to those of the system it replaces.

**Information More Readily Obtained From NHST**

There are actually two respects in which the p value associated with a NHST provides useful information that is not easily gleaned from the corresponding confidence interval: degree of confidence that we have not made a Type III error and likelihood that our sample result is replicable.

1. **The p value is approximately twice the probability of a Type III error.** The probability of a Type III error (rejecting $H_0$ in the wrong direction) equals α/2 when $H_0$ is true and drops precipitously (for two-tailed and split-tailed tests) as the true value of our parameter departs from its null-hypothesized value. Thus, although a highly significant (e.g., $p < .001$) result does not necessarily imply that the effect tested is large or important, it does indicate that you are considerably less likely to be wrong in your conclusion about the sign of the population effect than if your significance test had yielded a p value of, say, .049.

2. **A p value of .005 (note the extra zero) means the probability of exact replication is .80.** One of the major misinterpretations of significance tests is that a p value of .05 indicates that there is a 95% chance that a result would replicate (again be statistically significant in the same direction). In fact, the probability that a result that achieves a p value of exactly .05 in one study comes out statistically significant at the .05 level (in the same direction) in an exact replication is very close to .50. However, a p value of .005 (note the extra zero) is associated with about an 80% chance of a successful "exact" replication (Greenwald, Gonzalez, Harris, & Guthrie, 1996).

**NHST ALTERNATIVES’ SUSCEPTIBILITY TO MISINTERPRETATION**

We can examine the comments that have been made by ban proponents to see whether the proposed significant-test-free world is indeed likely to be less susceptible to misinterpretation than the current NHST-bedevedile world. Let us start with what Schmidt (1996) regards as the most pernicious misinterpretation of significance tests, namely, the tendency to interpret nonsignificance as establishing the validity of the null hypothesis. ("This belief has probably done more than any of the other false beliefs about significance testing to retard the growth of cumulative knowledge in psychology"); Schmidt, 1996, p. 126.) Schmidt inveighed against authors who summarize nonsignificant findings by saying that there was "no effect" of a given manipulation, solely on the basis of a lack of statistical significance, and pointed out that the resulting tabulation of studies that have found "no effect" versus those that have yielded significance leads to "wasted research efforts to identify non-existent [italics added] moderator variables" (p. 118). Similarly, Hunter (this issue) claims that each review based on such a search for moderator variables "can delay progress in an area of research for decades"—and that in personnel selection, "the use of the significance test has caused a 50-year delay in progression in some research areas!" But in declaring moderator effects nonexistent, Schmidt and Hunter have made exactly the same error of affirming the null hypothesis with respect to interaction effects that they excoriate with respect to main effects. Interactions are effects, too, and they are almost certainly not precisely zero.

Of even greater danger to the unwary reader are statements that appear to declare meta-analytic results impervious to sampling error. Thus, for instance, Schmidt (1996) stated that in meta-analysis, "[only] effect sizes are used, and significance tests are not used in analyzing the effect sizes" (p. 119), and that "[meta-analysis] tells us that there is only one population correlation, and that that value is .22" (p. 122).

Yes, meta-analyses will yield lower standard errors than the individual studies that form their database, but the standard errors will still be nonzero. We need a confidence interval around that .22.

Banning significance tests is clearly not going to guarantee that misunderstandings of the role of sampling error and the strength of evidence required to establish the validity of a null hypothesis will not continue to bedevil us.

**CONTROLLING DATA-CENSORSHIP BIAS AND THE TYPE I ERROR RATE**

Schmidt (1996) pointed out that precollection censoring of data leads to a more meager (or even nonexistent) database for any subsequent meta-analysis of the effect in question, whereas postcollection censoring of data by setting statistical significance as a necessary condition for publication leads not only to a smaller corpus of data but also to a positively biased estimate of the true population effect size.

The proposed alternative to NHST is that researchers be encouraged to carry out well-designed studies of any effects of potential importance without regard to considerations of (possibly very low) power, and that editors be asked to publish the resulting confidence intervals and point estimates without regard to statistical significance, power, or precision of estimate. Effectively, we would be resetting our α to 1.0. Ban proponents accompany this permissiveness with respect to publication of data, however, with an essentially total ban on interpretation of single-study results, deferring any conclusions about effect size and direction to the subsequent meta-analysis in which the uncensored data will be incorporated.

It seems unrealistic to expect authors and editors to completely avoid interpretation of results of single studies. A compromise would be to provide a medium for publication of nonsignificant results (presumably sheared of abstracts and Discussion sections that discuss the effects as if their signs had been definitely established), while reserving publication in archival journals for articles that can reasonably claim to provide convincing evidence as to sign of effect. There are now mass-storage systems with sufficient capacity to handle the deluge. However, someone would have to work out a good enough cataloguing system (key words, etc.) so that all studies relevant to a given effect could be retrieved by the meta-analyst.

**SUMMARY**

I remain optimistic that converting our presentations of NHST to three-valued logic will make NHST more consistent with sound research practice, will thus avoid many of the misinterpretations of NHST that have arisen over the decades, and will make clearer that any replacement of NHST will need to retain its element of social and
self-control over our tendency to overinterpret nearly random aspects of our data. I am also hopeful that the simpler procedures for introducing power considerations into our research planning that the MIDS and FEDS criteria (Harris & Quade, 1992; Harris, 1996) provide will help reform this aspect of NHST, whether it is retained in its present form or incorporated into reports of confidence intervals. Finally, I acknowledge that making statistical significance a requirement for publication of results leads to a positive bias in subsequent meta-analytic results. My preferred solution would be to impose statistical significance (or a zero-excluding confidence interval) as a condition for full-blown interpretation of one’s results, but to provide a means of archiving for subsequent use in meta-analyses all methodologically sound studies of a given effect, whether statistically significant or not.

1. A more detailed presentation of these and other arguments can be found in Harris (in press).

REFERENCES


Harris, R.J. (in press). Reforming significance testing via three-valued logic. In L. Harlow & S. Mulaik (Eds.), What if there were no significance tests? Mahwah, NJ: Erlbaum.


This document is a scanned copy of a printed document. No warranty is given about the accuracy of the copy. Users should refer to the original published version of the material.