When the Use of $p$ Values Actually Makes Some Sense

Barry H. Cohen
New York University


Abstract
The consequences of using NHST vary for different areas of psychological research, depending principally on the sizes of the samples typically employed, and on the sizes of the population effects being tested. In the case of mainstream academic experimentation, sample sizes rarely exceed 200 per condition, and it is reasonable to assume that no more than 20% of the experiments involve true population effects as tiny as, say, .01. Under those constraints, a significant result implies not only that the sample effect size is no less than “small” (i.e., Cohen’s $d = .2$), but that the population effect is unlikely to be tiny. In other words, the common fallacy, that a statistically significant result implies a population effect size large enough to be worthy of further and more specific experimentation, is approximately true within many areas of psychological research.

Introduction
The seventy-year-long, voluminous literature critical of null hypothesis significance testing (NHST) has been surprisingly ineffective in reducing the use of this statistical procedure in mainstream psychological research (Cumming, et al., 2007). For those who had hoped that the sixth and latest edition of APA’s publication manual, published just one month ago, would help to discourage the use of NHST in the future, the following passage must come as a disappointment: “Historically, researchers in psychology have relied heavily on … (NHST) as a starting point for many (but not all) of its analytic approaches. APA stresses that NHST is but a starting point and that additional reporting elements such as effect sizes, confidence intervals, and extensive description are needed to convey the most complete meaning of the results. The degree to which any journal emphasizes (or de-emphasizes) NHST is a decision of the individual editor.” (p. 33). However, we know from the survey I just referred to, conducted by Professors Cumming and Fidler and their colleagues, that among editors of some of the most prominent psychology journals “there were few indications of intentions to seek changes to statistical practices.” (p. 231). I will argue in this paper that editors have good reasons to cling to NHST, whether or not they can articulate, or are even aware of, those reasons.

Lest it seem that I am setting up a straw man, in that no one in the field is seriously calling for the end of the practice of NHST, I offer the following quote from Kline (2004), in a book, Beyond Significance Testing, recently published by the APA, and touted in its preface as “a follow-up report to the report of Leland Wilkinson and the Task Force on Statistical Inference …” (p. xi). Near the beginning of his third chapter, Kline gives us a preview of his position by stating that: “After review of the debate about NHST, I argue that the criticisms have sufficient merit to support the minimization or elimination of NHST in the behavioral sciences” (pps. 61-62). Bear in mind that in a footnote in the latest edition of APA’s Publication Manual, the reader is referred to Kline (2004) as one of four references for those interested in finding out more about the NHST controversy.

Perhaps, one reason that critics of NHST have failed to influence a large proportion of psychologists is that their arguments have been too sweeping in their generalities, and have not addressed the consequences of NHST within any particular areas of psychological research. It is clear to me that the various criticisms of NHST do not apply equally well across the many types of research performed by psychologists. I will demonstrate this point by illuminating the real-world consequences
of the widespread use of NHST within any subarea of psychology that is dominated by a form of research that I call the Academic Theory Corroboration Experiment (ATCE). I define an ATCE research area as one in which the ATCEs are bound by the following constraints: 1) No comparison or contrast of group means involves more than a total of 400 participants; 2) No more than 20% of the studies in this domain involve tests of a true null or even near-null hypothesis; 3) A maximum value for alpha is fixed rather rigidly by convention. In psychology, the use of the .05 level as the default value for alpha is extremely common. Results corresponding to p values between .05 and .1 are generally not described as significant, but rather as approaching, or exhibiting a trend towards, significance (but not unless the authors were hoping for significance). Conversely, declaring that “all p’s > .11,” is often used as a way of conveying that a set of results (perhaps, regarding potential confounding variables) is clearly not significant. Of course, smaller alpha levels are routinely used when conducting multiple tests within the same study.

The term “theory corroboration” indicates that the scale used to measure the dependent variable may have no practical meaning outside of the study itself – that is, a significant difference on the DV in the expected direction can be consistent with, and therefore supportive of, a psychological theory, but the amount of the difference has no physical or standard interpretation. Therefore, the use of confidence intervals (CI) for the dependent variable is generally not very informative in this domain, and even CI’s for a standardized effect-size measure may be of relatively little interest – though I hasten to add that I would never want to discourage the reporting of these supplements to NHST. A typical example of research that falls within the ATCE domain is a social psychology experiment that involves the random assignment of undergraduate psychology students to various experimental conditions or combinations of conditions, and measures, as its dependent variable, various attitudes by responses on a Likert scale. If at any point you feel that you can disregard my arguments about NHST, because you believe that ATCE’s are not very common in psychological research, please make a mental note to contact me later for a list of well-respected psychology journals that will change your mind about the prevalence of ATCEs.

An Immediate Consequence of Using NHST with ATCEs

A common criticism of NHST is that the label “statistical significance” is uninformative, because even the tiniest of population effect sizes (PES), other than zero, can be made likely to yield significant results, if it is investigated with large enough samples. However, this criticism does not apply to ATCEs, as I have defined them. Consider the comparison of two samples in which neither sample contains more than 200 observations. If the two samples are the same size, the t value for this comparison can be expressed as \( t = g \sqrt{\frac{n}{2}} \), where n is the size of each sample, and g is the standardized sample effect size as defined by Hedges (1981), and often referred to as Cohen’s d. At the maximum sample size stated for ATCEs, the two-group t test formula can be simplified to \( t = g \sqrt{\frac{200}{2}} \) = 10g. If we know that a t value is statistically significant at the .05 level with a two-tailed test, we know that the observed value must be at least 1.96, and therefore we also know that 10g > 1.96, so that the observed effect size must be at least slightly greater than 1.96/10 = .196, which is very close to the effect size that Cohen (1988) defined as being small (i.e., .2). So, to summarize, if we know that a t test of two independent sample means is significant at the .05 level (two-tailed), and we know that the total N will never exceed 400, we automatically know that the observed effect size cannot be less than the amount that Cohen referred to as “small.” Even without being aware of the underlying math, those working in research areas for which any one group or cell very rarely contains more than 200 observations, may be in the habit of assuming that any significant result they find will be associated at
least with an observed effect size that is not less than what is conventionally known as a small effect, and that assumption would be a reasonable one.

Does Sample Effect Size Tell Us Anything about Population Effect Size?
Although it is reasonable to assert that any particular near-zero PES is not very likely to produce a sample effect size large enough to yield a significant result with limited sample sizes, this does not tell us what we really want and need to know. The main problem with NHST is that finding a large and statistically significant effect in our data does not, by itself, allow us to conclude that the underlying PES is very unlikely to be tiny, or even zero. This can be called the base rate or “Bayesian” problem. That is why I added a second constraint to the ATCE domain – that no more than 20% of PES’s tested are at or very near zero (say, less than .01). Now let us look at the implications of obtaining statistically significant results, assuming the maximum value of this second constraint. If 10,000 experiments are conducted at the .05 level (two-tailed) with limited sample sizes, we can expect that, of the 2,000 near-zero PES’s, only about 100 will attain significance, and these can be considered either literal Type I errors, or at least grossly overestimated and misleading results (e.g., the test is significant, but the PES is only about .01). On the other hand, if the power of the other 8,000 studies is about 40%, as is often conservatively estimated for psychology experiments, then about 3,200 of them will attain significance. So, looking at the total of 3,300 significant results, we know that only 100 of them, or about 3%, will be Type I or Type I-ish errors. In the other 97% of the significant results, the PES will not be very small. So, with these constraints, the common fallacy – that about 5% of all results significant at the .05 level are really Type I errors – is approximately true, though a bit exaggerated. Note that in a research domain with the opposite base rates – 80% of PES’s tested are around zero (e.g., only alternative medical remedies are being tested) – but the same sample sizes, alpha, and power estimates apply, the percentage of significant results corresponding to near-zero PES’s will be about 33%. Fortunately, I think it is safe to declare that the base rates for most of psychological research resemble the former example much more closely than the latter.

Within the ATCE domain, NHST serves as an effective filter against tiny PES’s; only about 5% of the tiny PES’s that are tested will lead to significant results, and as long as the vast majority of PES’s being tested are not tiny, less than 5% of significant results will have arisen from tiny PES’s. (You can see from Figure 1 that, when dealing with a PES as small as .01, power hardly changes at all with sample sizes that never exceed 200 per group – and even with a PES five times larger than .01, the change in power is quite small within the sample-size constraints of the ATCE domain.)

Is the Null Hypothesis Ever Really True For Psychology Experiments, and Does it Really Matter?
So far, I have been treating true-null and near-null PES’s as equivalent, thus dodging the issue of whether a point null hypothesis is ever, or can ever be, exactly true in psychological research. It may seem that this is a serious neglect on my part, particularly given that the putative purpose of NHST is to give us some basis upon which to judge whether a point null hypothesis may be true or not. However, within the ATCE realm, exact null hypotheses are hardly ever of interest. If a particular causal mechanism or pathway is thought not to exist at all (e.g., conscious control of autonomic bodily functions), then merely rejecting the null hypothesis with some confidence, no matter how tiny the true effect size, can be of great interest. But such cases are rare in psychological research. Although the language of results sections in published articles makes it appear that the primary concern of the authors is whether the null hypothesis is literally true or not, in the vast majority of cases, NHST is really being used to establish whether the PES is of a tiny size or not; the researchers in such cases would not be comforted to know that the PES underlying their study is not zero, if it might well be extremely tiny.
It has been proposed that point nulls in psychology be replaced by tiny intervals, and in particular, that “nil” hypotheses be replaced by tiny intervals surrounding zero, such that the size of the interval is comparable to the error involved in measuring the DV (e.g., the good-enough ranges proposed by Serlin & Lapsley, 1993). Setting aside the potential difficulty of discerning the proper ranges for every DV used in psychology (and gaining consensus on these sizes), I would argue that given the power of a typical psychology experiment there is virtually no practical difference between testing an exact point null and a tiny range null that is comparable to measurement error. Only in research areas where very large samples are often used would this distinction become relevant. At the maximum of 200 observations for each of two samples, the power to detect an effect size of .01 differs very little from the power for an effect size of .001 – both are very close to the alpha level being used – and even an effect size of .1 yields a value for power that is only about .17 (see Figure 1).

**Is Screening Out Tiny Effect Sizes a Good Thing?**

Within the ATCE domain, NHST is highly effective at screening out results obtained from testing tiny PES’s in that the observed results rarely reach conventional significance levels, and therefore are rarely submitted for publication, and tend not to be published in the more selective journals even if they are submitted. However, as Prentice and Miller (1992) pointed out, the size of a population effect depends on the magnitude of the manipulation applied to the IV, as well as the sensitivity of the DV being measured. Surely, not all tiny PES’s are trivial or uninteresting, and some may be quite impressive, given the nature of the manipulations used. That is why it is so important to stress the original approach of Fisher (1955) with respect to cases in which the null hypothesis cannot be rejected, and insist that failures to reject not be used indiscriminately as evidence that the PES is zero. Nonetheless, even given this caveat it can be reasonably argued that the screening out of nearly all tiny PES’s, regardless of the possible importance or impact of the effect, is a fatal drawback of NHST in the ATCE domain, and reason enough to end this practice for such research. However, whether this flaw is truly a fatal one can only be decided with respect to comparing NHST to some alternative system for deciding which studies will be submitted for publication, and which will be published after being submitted. Before I consider any alternatives though, note that I am including under the umbrella of NHST any data analytic system that yields a probability for obtaining data as extreme as your own when the appropriate “nil” (i.e., no effect) hypothesis is true, and enforces the use of a specific maximal value for alpha with respect to labeling your results as significant or not. Therefore, any system that seeks only to improve the accuracy of the p values obtained (e.g., robust statistics, resampling methods), or to re-express those p values in more informative ways (e.g., $p_{rep}$) is still a form of NHST and not an alternative to it. Moreover, to the extent that 95% confidence intervals are used to determine statistical significance (e.g., by seeing if the nil hypothesis of zero is contained in the interval), these CI’s should be considered supplements to, rather than replacements of, NHST. As a true alternative to NHST, I will next consider the consequences for the publication of results from ATCEs when p values are either not obtained at all, or they are not compared to a standard alpha level in order to make a decision about them or label them in some way.

**Publication Without NHST**

Although the latest edition of APA’s Publication Manual urges the inclusion of effect sizes and CIs as supplements to, rather than replacements for, NHST, there are certainly statistical reformers out there who would like to see journal editors encouraging authors to submit any results exhibiting a predicted or interpretable pattern, along with the appropriate descriptive statistics, effect-size measures, and/or corresponding confidence intervals, and without p values and the language associated with NHST.
So, to gain some perspective on the role that NHST can play in psychological research, let us consider the consequences for the publication of ATCE results if the use of NHST and its associated \( p \) values were to be strongly discouraged.

First, it is easy to envision a very large increase in the number of MS’s submitted to ATCE journals, as researchers would be less likely to relegate to their file drawer studies in which the principal results do not attain conventional levels of statistical significance. Second, reviewers and editors would then be faced not only with more MS’s to evaluate, but with a new dimension on which to base publication decisions. They would have to decide how much weight to give to the sizes of estimated population effects among studies of equally sound design and theoretical import. This becomes an important consideration in that some of the submitted studies might present small observed effect sizes as evidence of a population effect in a particular direction. Unfortunately, any two-group comparison that is underpowered enough to be expected to yield a small \( t \) value, would be associated with a considerable probability of a Type III error (formerly known as the “error of the third kind”). If a Type III error is defined broadly, in the two-group case, as any observed result in which the order of the sample means is opposite to the order of the true population means, a result that would have yielded a \( t \) value of 1.0 (e.g., an effect size of .2 with 50 participants in each sample) had a \( t \) test been conducted would be associated with an expected Type III error rate of about .16. However, if we use NHST to define a Type III error as a statistically significant result in the wrong direction, the Type III error rate for an expected \( t \) value of 1.0 reduces to .0013.

**Getting Real about NHST**

Finding that a result is statistically significant does not really tell us anything about nature or the universe that we didn’t know just from our descriptive statistics. It certainly does not tell us that there is a “significant” effect in the population, whatever that might mean. Nor does finding that a result is not statistically significant tell us that there is no effect in the population, or even that whatever effect there is is too small to be concerned about. Indeed, it should be noted, as it has been many times by critics of NHST, that results which yield a \( p \) value of .04 are telling us a story that is very similar to the one told by results that lead to a \( p \) of .06. So, what does NHST really tell us? That is, what function can possibly be served by making an all-or-nothing decision that an obtained result is either significant or not? The answer is that, within a subfield of psychology in which researchers use fairly similar sample sizes, similar experimental designs (e.g., using repeated measures can be like using larger samples), and the same maximum value for alpha (normally, .05), NHST serves as a crude and rather arbitrary way to decide when a sample effect size is large enough to be worth telling others about – i.e., publishing. I cannot prove it, but I believe that the reason almost no one wants to confront the truth about NHST is that coming to any explicit agreement directly about effect sizes – e.g., how small does an effect have to be to be considered trivial or worth ignoring? – at this late stage in the development of psychological research, would be virtually impossible, even within a narrow subarea of psychology, and that reviewers, editors, and researchers all know this intuitively, even though few of them could articulate the basis of their loyalty to NHST.

The fact that researchers perform NHST themselves, and then refrain from submitting nonsignificant results, is an enormous convenience to the whole system of peer review and publication within psychology. But surely that convenience cannot justify the widespread use of NHST, with all the confusion that this widely misunderstood procedure has caused and continues to cause. Or can it? Frank Schmidt, a former president of Division 5, famously posed this challenge back in 1996: “Can you articulate even one legitimate contribution that significance testing has made (or makes) to the research enterprise (i.e., any way in which it contributes to the development of cumulative scientific
knowledge)?" My answer to this challenge is the following: The necessity to attain statistical significance for publication has pushed many a researcher to use larger samples than he or she would otherwise be inclined to use; thus fewer studies are conducted, but the ones that are conducted yield more reliable and precise results (e.g., narrower confidence intervals for population parameters, whether or not these are actually reported), with a smaller chance of committing directional errors.

Conclusions and a Call to Action

So, am I arguing in favor of the status quo? Am I trying to encourage researchers to keep using NHST the same way they have always done? Certainly not. My chief purpose in writing and presenting this paper was ultimately to improve statistical education. At present, the way NHST is explained in statistics textbooks is not very compelling. For example, as part of the rationale behind NHST, students are taught that when a “null” experiment has less than a one in twenty chance of producing results as extreme as the one you obtained (i.e., \( p < .05 \)), the null hypothesis becomes such an unlikely explanation for your results that it becomes safe to ignore. Really? Even if one knows nothing of the base rate for true nulls or “near nulls,” one out of twenty is not an impressively low probability, especially when considering the sheer number of psychological studies being conducted these days. And, especially when compared to the very tiny \( p \) values routinely obtained from DNA evidence in criminal trials, with which one can feel reasonably safe in rejecting the null hypothesis, while ignoring the base rate problem at the same time. Of course, if a large proportion of statistical tests involved true nulls, the popular alpha level of .05 would soon become unacceptably high.

Will there come a day when psychology students are taught that NHST by itself tells you absolutely nothing about the probability that the null (or even a near-null) is true for your experiment, that statistically significant results are not really “special,” and that researchers in many subareas of psychology are just following an arbitrary, somewhat haphazard, convention that they don’t fully understand, in order to reduce the number of studies submitted for publication for which the observed effect sizes are fairly small, and quite possibly in the wrong direction? It seems unlikely, given how embarrassing the story sounds when stripped of its pseudo-scientific formalities. However, in the meantime, it would not be a bad idea to correct the most egregious errors in the textbooks we use to train the next generation of psychologists.

For example, five years ago I found the following erroneous statement in the most popular college statistics text used at the time in psychology departments: “Consider, for example, a scientific journal containing 20 research articles, each evaluated with an alpha of .05. With a 5% risk of a Type I error, 1 out of 20 articles is probably a false report.” (p. 244). I reported this error, and a few others, to the senior author of the text, who graciously offered to make corrections as soon as possible. As both a textbook author and reviewer myself, I know firsthand how stats texts get published on the strength of a few sample chapters, and that few if any such texts are subjected to thorough fact-checking in their entirety either before or after publication. If the APA as an organization ever decides that improved statistical education of psychology students is an important priority for the future of psychology as a science, a task force should be formed to review all of the popular stats textbooks in the field thoroughly for erroneous and/or misleading statements, to recommend corrections to the authors as needed, and to award an APA seal of approval only to those texts that ultimately meet its standards of correctness. There is little point to making recommendations in a publication manual concerning the analysis and reporting of psychological data, if the textbooks students learn from do not help them to understand the justifications for those recommendations, and the reasoning behind them.

Before closing, I would like to make another plea to the membership of APA to take action – this time to mitigate perhaps the most damaging effect of the widespread use of NHST: the overestimation of
effect sizes in the published literature. In fact, even if NHST were banned completely, so long as the editors and reviewers of prominent journals were still to favor the publication of studies exhibiting larger effect sizes as compared to equally sound studies with smaller effect sizes, the effect sizes estimates found in the literature would be misleadingly large. Fortunately, the solution is now technologically quite simple. After a basic screening for sound methodology, all nonsignificant results could be posted on the web through one of APA’s existing mechanisms, in particular, PsychEXTRA. Submitted results would be required to follow a standard format which would ensure the inclusion of all the necessary data for subsequent meta-analyses, but would otherwise keep the reports brief. This database could be searched to see if an experimental manipulation has been tried and found to fail many times previously, or to conduct a power analysis for a prospective study, or a meta-analysis of results found in this database, with the proviso that the results posted in PsychEXTRA have not been subjected to the rigorous peer review of articles published in the more selective journals. Certainly it seems that, in the aggregate, a more accurate estimate of various effect sizes would emerge. Given how useful it would be to have access to virtually all of the studies attempted in psychology, whether successful or not, and given the enormous potential savings in time and money that could be gained from reducing duplication of efforts across the field, I was surprised at the total lack of enthusiasm that greeted my attempt to create a portal to PsychEXTRA that could be called PsychFILEDRAWER. Instead, I was offered vague critiques, such as: it will surely be too expensive; no one will bother submitting articles to such a website; and it will be hard to decide what to do about multiple studies and/or multiple findings within studies. And yet, at this very conference, X thousand posters were submitted for review, X hundred psychologists served as volunteer reviewers, and X hundred posters will be presented this weekend. So, I remain unconvinced that PsychFILEDRAWER is not worth doing, and I find it ironic that the official stance of APA has, in some ways, moved closer to discouraging NHST, while at the same time this very body, which is the only organization that could easily create and maintain PsychFILEDRAWER, seems to have so little interest in making nonsignificant results readily available to the entire field.

Before closing, I’d like to address the main question that many readers of this paper may have at this point: Is he for or against NHST? He writes disparagingly of NHST, and yet argues that it is useful. Well, my mission is to tell the truth and educate. Unfortunately, in the case of NHST, the truth is rather confusing and also embarrassing. As it stands now, the useful purpose that NHST serves in many areas of psychology, filtering out tiny effect sizes, does not match its rationale as described in statistics textbooks, and therefore few psychologists seem to know why they are using NHST, other than that it is required to get your results published in most journals. So, if my paper seems to be sending mixed messages, I would claim that it is just accurately reflecting a strange reality that now exists in many areas of psychological research – and clear and accurate communication on this topic is the best way to deal with that strangeness.

References


**Figure 1.**

![Power as a Function of Sample Size for Different Effect Sizes](image-url)