Human Communication Research ISSN 0360-3989

ORIGINAL ARTICLE

A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research

Timothy R. Levine[1], René Weber[2], Craig Hullett[3], Hee Sun Park[1], & Lisa L. Massi Lindsey[1]

1 Department of Communication, Michigan State University, East Lansing, MI 48823
2 Communication, University of California, Santa Barbara, CA 93106
3 Communication, University of Arizona, Tucson, AZ 85721

Null hypothesis significance testing (NHST) is the most widely accepted and frequently used approach to statistical inference in quantitative communication research. NHST, however, is highly controversial, and several serious problems with the approach have been identified. This paper reviews NHST and the controversy surrounding it. Commonly recognized problems include a sensitivity to sample size, the null is usually literally false, unacceptable Type II error rates, and misunderstanding and abuse. Problems associated with the conditional nature of NHST and the failure to distinguish statistical hypotheses from substantive hypotheses are emphasized. Recommended solutions and alternatives are addressed in a companion article.

doi:10.1111/j.1468-2958.2008.00317.x

> [Statistical significance testing] is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research.
>
> —Rozeboom (1960)

> Statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published.
>
> —Lykken (1968)

Corresponding author: Timothy R. Levine; e-mail: levinet@msu.edu
A version of this paper was presented at the Annual Meeting of the Internal Communication Association, May 2003, San Diego, CA. This paper is dedicated to the memory of John E. Hunter.

> I suggest to you that Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost exclusive reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories … is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology … I am not making some nit-picking statistician's correction. I am saying that the whole business is so radically defective as to be scientifically almost pointless.
>
> —Meehl (1978)

> Testing for statistical significance continues today not on its merits as a methodological tool but on the momentum of tradition. Rather than serving as a thinker's tool, it has become for some a clumsy substitute for thought, subverting what should be a contemplative exercise into an algorithm prone to error.
>
> —Rothman (1986)

> Logically and conceptually, the use of statistical significance testing in the analysis of research data has been thoroughly discredited.
>
> —Schmidt and Hunter (1997)

> Our unfortunate historical commitment to significance tests forces us to rephrase good questions in the negative, attempt to reject those nullities, and be left with nothing we can logically say about the questions.
>
> —Killeen (2005)

Null hypothesis significance testing (NHST) is the dominant approach to statistical inference in quantitative communication research. But as can be seen in the quotations above, NHST is also highly controversial, and there are many who believe that NHST is a deeply flawed method. Fisher (1925, 1935, 1995) and Neyman and Pearson (1933), who developed modern NHST, disagreed vehemently about how hypotheses should be tested statistically and developed statistical models that they believed were incompatible (Gigerenzer et al., 1989). A fusion of their ideas was introduced to the social sciences in the 1940s, and this so-called hybrid theory that has become modern NHST is institutionally accepted as the method of statistical inference in the social sciences (Gigerenzer et al., 1989). NHST is presented in social science methods texts "as the single solution to inductive inference" (Gigerenzer & Murray, 1987, p. 21), and commonly used statistical software packages such as SPSS and SAS employ NHST. Nevertheless, the use of NHST remains controversial and is often misunderstood and misused.

The use of NHST has been debated extensively in psychology (e.g., Bakan, 1966; Harlow, Mulaik, & Steiger, 1997; Nickerson, 2000; Rozeboom, 1960), education (e.g., Carver, 1978, 1993), sociology (e.g., Kish, 1959; Morrison & Henkel, 1970),

and elsewhere (e.g., Bellhouse, 1993; Berger & Sellke, 1987; Goodman, 1993; Rothman, 1986) for more than 40 years and is considered by its detractors as having been "thoroughly discredited" (Schmidt & Hunter, 1997, p. 37). Even the defenders of NHST acknowledge that it provides very limited information and that it is frequently misunderstood and misused (e.g., Abelson, 1997; Nickerson, 2000). In published communication research, however, NHST has been adopted with only superficial recognition of the problems inherent in the approach (cf. Chase & Simpson, 1979; Katzer & Sodt, 1973; Levine & Banas, 2002; Smith, Levine, Lachlan, & Fediuk, 2002; Steinfatt, 1990; for a notable exception, see Boster, 2002). Even in the best communication journals, misunderstanding and misinterpretation of NHST are the norm rather than the exception. Thus, communication research should benefit from a review of the controversy and recognition of the available alternatives.[1]

This paper offers a review and critique of NHST intended specifically for communication researchers. First, a brief history is provided and the NHST approach is described. Next, four common criticisms of NHST are reviewed. These are problems with the procedure's sensitivity to sample size, a concern that the null hypothesis is almost never literally true, concerns over statistical power and error rates, and concerns over misunderstanding about the meaning of statistical significance and corresponding misuse. Next, two more damaging but lesser known criticisms are summarized. These are the conditional nature of NHST and the problem of inverse probability and the failure to distinguish statistical hypotheses from substantive hypotheses. Alternatives and solutions are covered in a companion article (Levine, Weber, Hullett, & Park, 2008).

A review of NHST

A brief history

By one account, the first rudimentary significance test can be traced back to 1710 (Gigerenzer & Murray, 1987), but modern significance testing has developed since 1900. Karl Pearson, perhaps best known for the Pearson product–moment correlation, developed the first modern significance test (the chi-square goodness-of-fit test) in 1900 and soon after Gosset published work leading to the development of the *t* test (Student, 1908).

The two most influential approaches to modern NHST, however, were developed by Fisher (1925, 1935) and Neyman and Pearson (1933) in the early and mid-1900s (Gigerenzer & Murray, 1987; Kline, 2004). Fisher's approach to statistical hypothesis testing was developed as a general approach to scientific inference, whereas the Neyman–Pearson model was designed for applied decision making and quality control (Chow, 1996; Gigerenzer & Murray, 1987). In the Fisher approach, a nil–null hypothesis is specified and one tests the probability of the data under the null hypothesis. A nil–null hypothesis, often used in conjunction with a nondirectional alternative hypothesis (i.e., a two-tailed test), specifies no difference

or association (i.e., a nondirectional nil–H_0: effect = 0). Depending on the probability, one either rejects or fails to reject the null hypothesis (Fisher, 1995). The use of random assignment in experiments, null hypotheses, the analysis of variance, properties of estimators (i.e., consistency, efficiency, sufficiency), and the $p < .05$ criterion are some of Fisher's notable contributions (Dudycha & Dudycha, 1972; Gigerenzer et al., 1989; Yates, 1951). Perhaps Fisher's single most influential legacy, however, was his contention that NHST provides an objective and rigorous method of scientific inference suitable for testing a wide range of scientific hypotheses (Gigerenzer & Murray, 1987).[2] It was likely that this contention, along with the desirability of dichotomous reject–support outcomes, made significance testing appealing to social scientists (Gigerenzer & Murray, 1987; Krueger, 2001; Schmidt, 1996).

Neyman and Pearson (1933) offered what they believed to be a superior alternative to the Fisher approach. Unlike Fisher, the Neyman–Pearson approach specifies two hypotheses (H_0 and H_1) along with their sampling distributions. This provides for an estimation of Type II error and statistical power that are not defined in Fisher hypothesis testing (Gigerenzer & Murray, 1987). The Neyman–Pearson approach also requires that alpha is set in advance.

Heated debate between Neyman–Pearson and Fisher ensued. Both sides saw their models as superior and incompatible with the other's approach (Gigerenzer et al., 1989). This debate remains unresolved but has been mostly ignored in the social sciences where the two approaches have been fused into a widely accepted hybrid approach. As Gigerenzer et al. (1989) describe it:

> Although the debate continues among statisticians, it was silently resolved in the "cookbooks" written in the 1940s to the 1960s, largely by nonstatisticians, to teach in the social sciences the "rules of statistics." Fisher's theory of significance testing, which was historically first, was merged with concepts from the Neyman-Pearson theory and taught as "statistics" per se (p. 106). It is presented anonymously as statistical method, while unresolved controversial issues and alternative approaches to scientific inference are completely ignored (pp. 106–107). The hybrid theory was institutionalized by editors of major journals and in the university curricula (p. 107). As an apparently noncontroversial body of statistical knowledge, the hybrid theory has survived all attacks since its inception in the 1940s (p. 108). Its dominance permits the suppression of the hard questions (p. 108). What is most remarkable is the confidence within each social-science discipline that the standards of scientific demonstration have now been objectively and universally defined (p. 108).

A brief description of NHST

In modern hybrid NHST, there are two mutually exclusive and exhaustive statistical hypotheses, the null (H_0) and the alternative (H_1). The alternative hypothesis typically reflects a researcher's predictions and is usually stated in a manuscript. The null

hypothesis is the negation of the alternative hypothesis. For example, if a researcher predicts a difference between two means, the alternative hypothesis is that the two means are different and the null is that the means are exactly equal. The null hypothesis is seldom stated in research reports, but its existence is always implied in NHST.

The most common form of null hypothesis is a nil–null that specifies no difference, association, or effect and is associated with two-tailed tests (Nickerson, 2000). Alternatively, when one-tailed tests are used, the null hypothesis typically includes the nil–null and all other wrong direction findings (i.e., a directional nil–H_0: effect ≤ 0 or effect ≥ 0). Other types of null hypotheses are possible, such as in effect significance testing, but the nil–null and the nil plus-a-tail null are most common in communication research.

In standard hybrid NHST, a researcher selects a single arbitrary alpha level a priori, usually the conventional $\alpha = .05$. Once data are collected, a test statistic (e.g., t, F, χ^2) and a corresponding p value are calculated, most often by computer. The p value indicates the probability of obtaining a value of the test statistic that deviates as extremely (or more extremely) as it does from the null hypothesis prediction if the null hypothesis were true for the population from which the data were sampled. If the p value is less than or equal to the chosen alpha, then the null hypothesis is rejected on the grounds that the observed pattern of the data is sufficiently unlikely conditional on the null being true. That is, if the data are sufficiently improbable if the null were true, it is inferred that the null is likely false. Because the statistical null hypothesis and the statistical alternative hypothesis are written so that they are mutually exclusive and exhaustive, rejection of the null hypothesis provides the license to accept the alternative hypothesis reflecting the researcher's substantive prediction. If, however, the obtained p value is greater than alpha, the researcher fails to reject the null, and the data are considered inconclusive. Following Fisher (1995), null hypotheses are typically not accepted. Instead, one makes a binary decision to reject or fail to reject the null hypothesis based on the probability of the test statistic conditional on the null being true.[3]

The commonly asserted function of modern hybrid NHST is to provide an objective and reasonably accurate method of testing empirical hypotheses by ruling out chance (specifically sampling error) as an explanation for an observed difference or association (Abelson, 1997; Greenwald, Gonzalez, Harris, & Guthrie, 1996). Objectivity is claimed on the grounds that both the hypotheses and the alpha level are stated a priori and that significance rests on an observable outcome. Accuracy is claimed because a precise and conservative decision rule is used. Only results that could occur by chance 5% or less of the time (conditional on a true null) merit the label statistically significant. Finally, and most importantly, NHST purportedly provides social scientists with a method of distinguishing probabilistically true findings from those attributable to mere chance variation (Abelson, 1997; Kline, 2004). On the surface, then, NHST appears to have many attractive characteristics, and it seems to serve an important and needed function. On closer inspection, however, several problems with the approach are evident.

Four common criticisms of NHST

Sensitivity to sample size

Perhaps the most widely recognized limitation in NHST is its sensitivity to sample size (e.g., Boster, 2002; Cohen, 1990). When the sample size is small, strong and important effects can be nonsignificant (i.e., a Type II error is made). Alternatively, when sample sizes are large, even trivial effects can have impressive-looking p values. For example, $r(20) = .40$ is not statistically significant at $p < .05$ (two tailed), whereas $r(1{,}000) = .07$ is statistically significant. Or, as another example, if an observed effect is exactly $r = .25$, the results are statistically significant if $n = 63$ but not if $n = 61$ (two tailed). As these examples demonstrate, the p values from null hypothesis significance tests reflect both the sample size and the magnitude of the effect observed, and obtaining or failing to obtain statistical significance is as much or more a function of one's sample size (and other things that affect statistical power such as measurement reliability, manipulation strength, meeting statistical assumptions, etc.) than the verisimilitude of one's substantive predictions or theory (Meehl, 1986). This can lead to dismissing potentially important findings when a sample is small and embracing trivial effects with large samples. There is merit to the argument that these are undesirable properties in a decision rule or a form of evidence and that NHST is therefore problematic.

As a consequence of the sample size problem, there is a growing recognition of the importance of reporting and interpreting effect sizes to supplement significance tests (e.g., Kirk, 1996). Various estimates of magnitude of effect or effect size tell us how strongly two or more variables are related or how large is the (mean) difference between groups. As Scarr (1997) observes, "perhaps the most egregious mistake is to confuse the statistical probability of an outcome with its theoretical or practical importance" (p. 17). Even though small effects are sometimes meaningful (Abelson, 1985), the theoretical and practical importance rest more on the magnitude of effect than on the probability of the data given the null hypothesis.

The point or nil–null is almost always false

A second common criticism of NHST is that a point or nil–null hypothesis is (at least in the social sciences) almost always literally false, independent of the verisimilitude of the substantive hypothesis (e.g., Cohen, 1994; Meehl, 1978).[4] Briefly, a substantive hypothesis refers to the knowledge claim underlying a research prediction, and substantive and statistical hypotheses are contrasted later in this essay. In any case, the observed correlation between any two variables, or the difference between any two means, will seldom be exactly 0.000 out to the nth decimal because there will always be uncontrolled spurious third variables in correlation-based studies or because randomization cannot be expected to exactly balance out the effects of all extraneous factors in experiments (Meehl, 1978).

If the null hypothesis is false anyway, then disproving it is both unimpressive and uninformative (Abelson, 1995; Cohen, 1994). When combined with the sample size

criticism, however, the "null is always false" argument has an even more troubling implication. If the null is always false, and if the sample size is large enough, a significance test will yield an affirmative result even if the researcher's substantive hypothesis is false (see Meehl, 1986). This means that substantive Type I error rates might be considerably higher than statistical Type I error rates and that nil–null NHST alone is not an accurate reflection of the verisimilitude of one's substantive hypothesis with large sample sizes.

Specifically, given that a statistical point null hypothesis is almost never literally true, statistical Type I errors should be infrequent and well below $p = .05$. This is because statistical Type I errors can only happen when the null is true (cf. Pollard & Richardson, 1987). So if alpha is set at $p = .05$, the expected frequency of Type I errors is 5% only when the probability of the statistical null is $p = 1.00$. As the probability that the null is actually true declines, so does the chance that any given NHST will yield a Type I error. As a consequence, if the statistical null was always literally false, a Type I error would be impossible. Hence, because the probability of the statistical null is typically considerably less than $p = 1.00$, the probability that an NHST will yield a Type I error under such conditions is considerable less than $p = .05$.

However, given that the long-term survival of theories in the social sciences used to generative substantive hypotheses is low (Meehl, 1978), it can be reasonably argued that substantive alternative hypotheses are substantially less probable than their statistical counterparts. Thus, substantive Type I errors are likely to occur at rates much higher than 5% and, consequently, statistical Type I error rates should be much lower than substantive Type I error rates. Simply put, just because a statistical null can be rejected does not mean that the corresponding substantive conclusions are correct and substantive false positives are almost certainly more prevalent in communication research than statistical false positives.

A related issue is Meehl's (1967, 1978, 1997) idea of risky tests. Meehl's argument rests on falsification (i.e., all hypotheses must be potentially falsifiable; Popper, 1959) and holds that tests of riskier predictions provide potentially stronger corroborative evidence than less risky tests (cf. Lakatos, 1978). That is, the greater the risk of falsification, the more impressive the results should the substantive predictions survive the test. Nil–null NHST provides only weak evidence for weak predictions because the predictions tested are highly imprecise (i.e., predicting, at best, direction of effect), and (assuming adequate statistical power) demonstrating statistically that the null hypothesis is unlikely is far from risky. It has been further argued that because the null hypothesis is never accepted in hybrid NHST, NHST is often open to criticism for a lack of falisifiability (Levine & Banas, 2002).

Power and error rates

A third criticism of NHST concerns unacceptably high Type II error rates in NHST as most often practiced (Boster, 2002; Hunter, 1997; Schmidt, 1996). This argument holds that analyses of social science research consistently show that published studies

lack adequate statistical power (e.g., Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Significance tests that lack power are relatively likely to produce Type II errors. Hunter (Schmidt & Hunter, 1997) notes that the statistical power in some literatures is so low that flipping a coin would provide a more accurate decision rule than NHST. Even in literatures where power is typically higher, Type II error rates make NHST an imprecise and error-fraught method that is ill suited to the needs of science (Smith et al., 2002).

Combined with "the null is always false" argument, the low-power argument gains even more force. When statistical power is low, Type II error rates make NHST highly problematic. As power increases, Type II errors are of less concern, but the "null is always false" argument becomes of greater concern. Therefore, researchers who rely on NHST to make dichotomous decisions about the viability of substantive hypotheses have an unacceptably high probability of drawing incorrect conclusions. A decision rule in which errors are probable and difficult to avoid cannot be a useful tool.

Power is a problematic issue in modern hybrid NHST for reasons other than just error rates. Specifically, power is most often undefined in modern NHST. As noted previously, power is a Neyman–Pearson idea, but modern NHST is a hybrid more heavily influenced by Fisher. The sampling distributions of both H_0 and H_1 are specified in Neyman–Pearson theory but not in Fisher or hybrid NHST. Instead, H_1 is simply specified to be not H_0 and vice versa. So, for example, if H_0: $r \geq .00$ and H_1: $r < .00$, power cannot be calculated for any sample size. Instead, something like H_0: $r \leq .00$ and H_1: $r \geq .20$, $n = 100$, that is, a specific alternative hypothesis, is required for power calculations. An effect size or point prediction must be specified for H_1 in order for power to be meaningful.

Misunderstanding and abuse

A final common criticism of NHST is that it is often misunderstood and abused. Carver (1978), Kline (2004), and Nickerson (2000) list many misconceptions about NHST. For example, the "odds-against-chance fantasy" is a label for the incorrect interpretation of a p value as the probability that a result is due to mere chance (Carver, 1978). Another example is the false belief that a $p < .05$ finding is one that is 95% likely to replicate (Carver, 1978; Greenwald et al., 1996; Killeen, 2005; Nickerson, 2000; Oakes, 1986). Yet another example of a false belief is that at $p < .05$, there is a 1 in 20 chance of making a Type I error (Pollard & Richardson, 1987).[5]

Perhaps the most common and most serious misunderstanding, however, is that a finding of $p < .05$ provides compelling probabilistic evidence in support of a substantive hypothesis (Greenwald et al., 1996; Kline, 2004; Meehl, 1978; Oakes, 1986; Trafimow, 2003). As shown below, a $p < .05$ finding does not mean that null hypotheses can be rejected with 95% or better confidence, and even if it did, the fact that the statistical null is false does not justify 95% or better confidence in a substantive alternative hypothesis (Kline, 2004). As the quotations from Rozeboom (1960) and Killeen (2005) provided at the beginning of this article mention,

concluding support for a substantive hypothesis on the basis of an NHST at $p < .05$ is not logically justified and instead rests on a misunderstanding of the meaning of statistical significance.

Summary

Common criticisms of NHST include a sensitivity to sample size, the argument that a nil–null hypothesis is always false, issues of statistical power and error rates, and allegations that NHST is frequently misunderstood and abused. Considered independently, each of these problems is at least somewhat fixable. For example, NHST could be interpreted in conjunction with estimates of effect size largely overcoming the first two criticisms above. Researchers could conduct power analyses in advance and collect sample sizes large enough (in conjunction with highly reliable measures, strong manipulations, robust applications, etc.) for acceptable Type II error rates. Education and standards could correct the problems of misunderstanding and abuse.

Three serious problems, however, remain. First, whereas defenders of NHST often argue that current practices should be blamed rather than the NHST procedure itself (Abelson, 1997; Frick, 1996; Hagen, 1997; Mulaik, Raju, & Harshman, 1997; Nickerson, 2000), it is clear that 40+ years of books and articles aimed at correcting the situation have largely failed and that these problems still exist.[6] Thus, to the extent that the problems reviewed above can be corrected, there is nevertheless ample evidence suggesting that the problems persist in spite of whether or not they are fixable.

Second, although the problems identified above are individually correctable, in combination, they are much more damaging and more difficult to correct. For example, consider the implications of Meehl's (1986, 1997) "crud factor" (also called "ambient correlational noise" by Lykken, 1968; see also Cohen, 1994; Oakes, 1986). The crud factor refers to possible systematic differences or associations between variables that exist and are observed due to systematic uncontrolled spurious third variables and extraneous factors and that are independent of the verisimilitude of the predicted relations. Thus, the crud factor contributes to the null almost always being false. The problem is, it is practically impossible to know what proportion of a significant effect might be due to crud and what is not. So even if sufficient statistical power exists, and even if effect sizes are taken into account, it is still difficult to interpret the results of an NHST with precision.

Third, two additional, and arguably more severe, problems with NHST exist. The first involves conditional and inverse probabilities and has been called the converse inequality argument (Markus, 2001). At issue is if a $p < .05$ result logically justifies rejecting H_0. It is argued that due to its conditional nature, NHST does not tell us the information we want to know (e.g., Carver, 1978; Cohen, 1994). The second issue is the extent to which rejecting a statistical H_0 provides logical support for a substantive H_1 (see, e.g., Meehl, 1986, 1997; Oakes, 1986). These two issues are reviewed in the following section.

Two criticisms of focus

Logical problems in rejecting H_0

A first criticism of focus stems from the conditional nature of p values in NHST and has been labeled the inverse probability error (Cohen, 1994) and the converse inequality argument (Markus, 2001). In standard hybrid NHST, the observed value of the test statistic is considered relative to a known probability distribution based on the null being true. Thus, NHST is informative about the probability of the data given the null but is used to draw an inference about the (im)probability of a null hypothesis given the data.

The probability of the data given the null hypothesis can be written as $P(D|H_0)$, whereas the probability of a null hypothesis given the data is $P(H_0|D)$. A serious problem for NHST is that $P(D|H_0) \neq P(H_0|D)$, and $P(H_0|D)$ cannot be determined based only on the knowledge of $P(D|H_0)$ (Cohen, 1994; Nickerson, 2000). Boster (2002) provides the example of the probability of a woman being a nurse versus the probability of a nurse being a woman. Clearly, these two probabilities are not the same. Because $P(D|H_0) \neq P(H_0|D)$, knowing that the data are 5% or less probable given the null does not mean the null is 5% or less likely given the data. NHST gives the inverse probability, not the probability of the null hypothesis or the probability of the null hypothesis given the data. Simply put, NHST does not tell researchers what they want to know, which is P (H|D) (Cohen, 1994; Kirk, 1996; Kline, 2004).

At least two potential (partial) solutions to the problem of inverse probability have been proposed. First, Bayes's theorem could be used to estimate $P(H_0|D)$ from $P(D|H_0)$ given a willingness to make subjective estimations of some other parameters. Nickerson (2000), for example, argues that by making some reasonable assumptions about the prior probability of the null and the probability of the data given the alternative, one might draw inferences about the probability of the null hypothesis based on p values. With this approach, Nickerson suggests that "as a general rule, a small p, say, $p < .001$, is reasonably strong evidence against H_0, but not as strong as usually assumed" (p. 252). Similarly, Trafimow (2003) shows that when the prior probability of the null is low and the probability of the finding given H_1 is high, small observed p values do provide quasi-accurate probabilistic evidence against a statistical null.

Second, the problem has been approached by considering the probability of replication. Although a p value should not be considered as informative about the exact probability of replication, Greenwald et al. (1996) estimated the probability of replicating findings and concluded that in many circumstances, a finding of $p \leq .005$ would have at least an 80% chance of replicating at $p < .05$. Taken together, these estimates suggest that small p values ($p < .005$ or $p < .001$) do constitute defensible (but qualified) evidence against a statistical null hypothesis with 80% or better confidence.

Nickerson's (2000), Trafimow's (2003), and Greenwald et al.'s (1996) demonstrations raise additional problems for those using NHST. First, it is clear that

findings of $p < .05$ do not logically or mathematically justify rejecting the statistical null hypotheses with 95% confidence and that a finding of $p < .05$ provides only weak evidence against a nil–null hypothesis. Second, whereas adopting an alpha of $p < .001$ might often allow for the desired 95% confidence in rejecting a statistical null hypothesis given certain assumptions, the practice of reducing alpha would dramatically lower statistical power, making the solution arguably worse than the problem. Third, as Trafimow points out, these issues pose an additional dilemma for NHST. For p values to be informative about the probability of the H_0, the prior probability of the null must be low. However, if the prior probability of the H_0 is low, then disproving the null provides little news value. As Trafimow puts it, "the valid performance of NHST implies little information gain, and gaining a lot of information implies an invalid use of NHST" (p. 533). Thus, there exist serious logical difficulties with rejecting a statistical null hypothesis based on p values from NHSTs, and the evidentiary value of $p < .05$ is meager. Moreover, attempts to increase the evidentiary value by lowering alpha would likely be counterproductive.

Logical problems in accepting substantive hypotheses

In the previous section, it was shown that NHST does not provide logical license to reject a statistical null hypothesis. In this section, it is shown that even if we could reject a null hypothesis on the basis of a small p value, making the inferential leap from rejecting the statistical null to accepting a substantive alternative hypothesis is questionable. One key issue here is the often ignored distinction between substantive and statistical hypotheses. The statistical alternative hypothesis (in Fisher testing) is the logical negation of the statistical null hypothesis, whereas a substantive hypothesis reflects the knowledge claim that the research is making (including substantive, conceptual, and theoretical meanings and implications). This distinction is essential when considering the scientific value of NHST (Edwards, 1965; Fowler, 1985; Kline, 2004; Meehl, 1997; Oakes, 1986). Researchers, of course, are interested in making substantive claims, and statistical analyses are only meaningful to the extent they are informative about the viability of substantive hypotheses.

In hybrid NHST, the statistical null hypothesis and the statistical alternative hypothesis are written such that they are mutually exclusive and exhaustive. If the statistical null is probably false, then the statistical alternative is inferred to be probably true on the grounds that no other alternatives (besides H_0 and H_1) are logically possible. Because it is possible for an alternative statistical hypothesis to be factually correct when the substantive thinking that gave rise to the hypothesis is false, probabilistic evidence for a statistical alternative hypothesis does not count as equally convincing evidence for the corresponding substantive hypothesis. That is, whereas the statistical null and the statistical alternative hypotheses are mutually exclusive and exhaustive by definition, such is not true for the statistical null and the substantive alternative. Multiple and conflicting substantive hypotheses can produce identical statistical alternative hypotheses, so accepting a statistical alternative hypothesis does not logically justify accepting a specific substantive hypothesis

consistent with the statistical alternative hypothesis. For example, both genetic- and environment-based theories predict that siblings will be more similar in communication traits than randomly paired others. Similarly, methodological artifacts can produce effects consistent with an alternative statistical hypothesis. As Trafimow (2003) notes, "it is absolutely crucial for researchers to derive hypotheses that are likely not to be true, absent the theory, if they want to have a chance at providing an impressive argument for that theory" (p. 533).

Accepting a substantive hypothesis on the basis of support for a statistical alternative hypothesis exploits the fallacy of affirming the consequent (Meehl, 1997). It is the case in a well-designed study that if the researcher's thinking is correct, then the statistical alternative hypothesis should be true. But it does not follow that if the statistical alternative hypothesis is true, a researcher's corresponding substantive hypothesis must be true. Thus, even if one can reject a statistical null, inferring support for a substantive alternative hypothesis on the basis of a false statistical null is problematic.

A final logical problem with accepting a substantive H_1 on the basis of a p value involves how probability is defined in NHST. The p values in NHST are based on a frequentist definition of probability that views probability as the relative frequency of occurrences in the long run (Gigerenzer & Murray, 1987; Gigerenzer et al., 1989; Oakes, 1986). Both Fisher's idea of a hypothetical infinite population and Neyman–Pearson repeated random sampling theory are based on this frequentist view of probability. A result with $p \leq .05$ is one that would occur 5% of the time or less conditional on the null being true. As Oakes (Meehl, 1997) points out, however, the probability of the truth or falsity of a scientific hypothesis makes little sense from a frequency view of probability. Unlike statistical hypotheses, substantive scientific hypotheses are not true or false some specifiable proportion of the time in the long run. Instead, an alternative view of probability is needed when considering evidence for or against a substantive hypothesis (Royall, 1997). Thus, NHST is based on a view of probability that is arguably incompatible with its application to scientific or substantive hypothesis testing.

Summary

As shown above, NHST, as used to assess the viability of social scientific hypotheses, involves a string of three inferences. Two of these are suspect. A statistical null hypothesis is either rejected or not depending on the probability of the data given the null. Rejecting the statistical null hypothesis leads to the acceptance of the statistical alternative hypothesis that, in turn, leads to the acceptance of a substantive alternative hypothesis. It has been shown here that neither rejecting the null based on the p value of an NHST nor accepting a substantive hypothesis based on acceptance of a statistical alternative hypotheses is logically or mathematically justified in modern hybrid NHST and that both of these inferences are tenuous. As a result, the evidence for a substantive hypothesis provided by a typical NHST is weak and nowhere near the 95% certainty often attributed to the test. It should now be clear

that, as noted in the opening quotations by Killeen (2005) and Rothman (1986), the continued use of classical NHST therefore rests on the power of convention and tradition rather than on its logical merit or its scientific utility.

Conclusions

This paper offers a critical evaluation of the use of NHST in communication research. As the opening quotations suggest, the practice is more controversial than most communication researchers probably realize. Although widely accepted and heavily used, serious problems with NHST exist, and these problems hinder scientific progress. The main thesis of the present paper is that awareness and understanding of these issues are essential for quantitative communication research practice and consumption.

Some problems with null hypothesis significance tests are likely well known. Most communication researchers, for example, understand that statistical significance is sample size dependent. Consistent with this, communication researchers, at least in the better journals, typically report estimates of effect size along with NHST. Many researchers also are likely familiar with the concept of statistical power, although many may underestimate the prevalence of Type II errors. Fewer researchers may be aware of the existence and implications of "the null is always false" argument and the fact that one can test many other (and better) null hypotheses than the nil–null hypothesis, which is seldom of real interest.

The criticisms of focus here are likely less well known and recognized but are more fundamental. Because the p values associated with NHST reflect conditional probabilities, one cannot logically reject a statistical null hypothesis with 95% confidence based on $p = .05$. Furthermore, due to the distinction between substantive and statistical hypotheses, rejection of the null does not logically justify acceptance of a substantive alternative hypothesis. Together, these severely limit the utility of NHST in assessing the verisimilitude of communication theories and research hypotheses. As the beginning quotations assert, these pose serious inferential obstacles that limit knowledge generation and scientific progress.

Fortunately, these problems can be minimized or overcome, and a number of solutions and alternatives are available. Supplementing significance tests with estimates of effect sizes, risky effect null hypotheses, confidence intervals, and a priori statistical power can counteract common problems associated with NHST. Increased reliance on descriptive statistics and meta-analysis would improve the state of knowledge. A gradual move toward riskier tests would allow research communities to better assess the viability of well-articulated theory. Most of all, a better awareness of the limitations of NHST is needed. Hopefully, this essay will provide communication researchers with an understanding of the issues involved in the long-standing but little recognized NHST controversy and, in conjunction with our companion article (Levine et al., 2008), suggest practices that will further scientific progress.

Notes

1 The commentary offered here provides only a brief and incomplete summary of NHST and the controversy surrounding it. Several readings are recommended to those who are interested in learning more about NHST. Informative histories of the development of NHST are provided in Cowles and Davis (1982), Dudycha and Dudycha (1972), Chapters 1–3 of Gigerenzer et al. (1989), and the first chapter of Gigerenzer and Murray (1987). Cohen (1990, 1994), Hunter (1997), Meehl (1967, 1978, 1986, 1997), Oakes (1986), and Schmidt (1996) provide different and influential critiques of NHST, and Abelson (1997), Chow (1996), Frick (1996), Hagen (1997), and Mulaik et al. (1997) offer defenses. Markus (2001) provides a detailed analysis of logical arguments made for and against NHST, and Kline (2004) and Nickerson (2000) provide recent reviews of the controversy.

2 Fisher suggested his idea of NHST as an objective methodology for experiments in fundamental research where little was known about potential effects. In this context, it was first of all necessary to test whether or not there is an effect. In other words, a pure "sign hypothesis" that tests the mere existence of an effect was the focus of his early work. In this context, Fisher's NHST is acceptable and still valuable. Today, however, research is more advanced and better null hypotheses can be offered, for example, H_0: effect $\leq \Delta$ (effect null hypotheses). Communication researchers, however, rarely do this, and this is one of our main critiques. One of the reasons is the necessity to use non-central sampling distributions. This issue is addressed in the companion paper.

3 The thoughtful reader will find this line of argument fallacious. It, of course, is. The specific logical flaws with NHST are detailed later in the article. The reasoning presented here is, nevertheless, to the best of the current authors' knowledge, an accurate portrayal of the (il)logic of nil–null NHST. It should also be noted that although not a specific point of focus in the current paper, the use of NHST to justify dichotomous reject versus fail-to-reject decisions has been criticized (e.g., Folger, 1989; Kirk, 1996; Rozeboom, 1960).

4 A corollary of this observation is that the probability that a nil plus-a-tail null hypothesis (i.e., with a one-tailed or directional hypothesis) is false is approximately 50% of the time, independent of the verisimilitude of the substantive hypothesis (Meehl, 1986; Oakes, 1986).

5 Each of these misconceptions rests on the fact that the p value obtained from a significance test is conditional on the null being true. This issue is explained in the section covering the first criticism of focus.

6 One exception to this is that recognition and reporting of effect sizes have improved considerably in published communication research. In this regard, communication research appears ahead of many other quantitative social sciences.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, **97**, 129–133.

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117–141). Mahwah, NJ: LEA.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, **66**, 432–437.

Bellhouse, D. R. (1993). P values, hypothesis tests, and likelihood. *American Journal of Epidemiology*, **137**, 497–498.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, **82**, 112–122.

Boster, F. J. (2002). On making progress in communication science. *Human Communication Research*, **28**, 473–490.

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

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, **61**, 287–292.

Chase, L. J., & Simpson, T. J. (1979). Significance and substance: An examination of experimental effects. *Human Communication Research*, **5**, 351–354.

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. London: Sage.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, **65**, 145–153.

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.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, **37**, 553–558.

Dudycha, A. L., & Dudycha, L. W. (1972). Behavioral statistics: An historical perspective. In R. E. Kirk (Ed.), *Statistical issues: A reader for the behavioral sciences*. Monterey, CA: Brooks Cole.

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, **63**, 400–402.

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

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1995). *Statistical methods, experimental design, and scientific inference: A re-issue of statistical methods for research workers, the design of experiments, and statistical methods and scientific inference*. Oxford, UK: Oxford University Press.

Folger, R. (1989). Significance tests and the duplicity of binary decisions. *Psychological Bulletin*, **106**, 155–160.

Fowler, R. L. (1985). Testing for substantive significance in applied research by specifying nonzero effect null hypotheses. *Journal of Applied Psychology*, **70**, 215–218.

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

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: LEA.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance: How probability changed science and everyday life*. New York: Cambridge University Press.

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

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**, 175–186.

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. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: LEA.

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

Katzer, J., & Sodt, J. (1973). An analysis of the use of statistical testing in communication research. *Journal of Communication*, **23**, 251–265.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, **16**, 345–353.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, **56**, 746–759.

Kish, L. (1959). Some statistical problems in research design. *American Sociological Review*, **24**, 328–338.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, **56**, 16–26.

Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers volume 1*. Cambridge, UK: Cambridge University Press.

Levine, T. R., & Banas, J. (2002). One-tail *F*-tests in communication research. *Communication Monographs*, **69**, 132–143.

Levine, T. R., Weber, R., Hullett, C. R., & Park, H. S. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research*, **34**, 188–209.

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

Markus, K. A. (2001). The converse inequality argument against tests of statistical significance. *Psychological Methods*, **6**, 147–160.

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

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, **46**, 806–834.

Meehl, P. E. (1986). What social scientists don't understand. In D. W. Fiske & R. A. Shweder (Eds.), *Meta-theory in social science* (pp. 315–338). Chicago: University of Chicago Press.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Mahwah, NJ: LEA.

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

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–115). Mahwah, NJ: LEA.

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient test of statistical hypothesis. *Philosophical Transaction of the Royal Society of London—Series A*, **231**, 289–337.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, **5**, 241–301.

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

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

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Routledge.

Rothman, K. J. (1986). Significance questing. *Annals of Internal Medicine*, **105**, 445–447.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. Boca Raton, FL: Chapman and Hall.

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

Scarr, S. (1997). Rules of evidence: A larger context for the statistical debate. *Psychological Science*, **8**, 16–17.

Schmidt, F. L. (1996). Statistical significance testing and the cumulative knowledge in psychology: Implications for training researchers. *Psychological Methods*, **1**, 115–129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: LEA.

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

Smith, R. A., Levine, T. R., Lachlan, K. A., & Fediuk, T. A. (2002). The high cost of complexity in experimental design and data analysis: Type I and type II error rates in multiway anova. *Human Communication Research*, **28**, 515–530.

Steinfatt, T. M. (1990). Ritual versus logic in significance testing in communication research. *Communication Research Reports*, **7**, 90–93.

Student [Gosset, W. S.]. (1908). The probable error of the mean. *Biometrika*, **6**, 1–15.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review*, **110**, 526–535.

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, **46**, 19–34.

Une évaluation critique du test de signification basé sur l'hypothèse nulle dans la recherche quantitative en communication

Timothy R. Levine

René Weber

Craig Hullett

Hee Sun Park

Lisa L. Massi Lindsey

Résumé

Le test de signification basé sur l'hypothèse nulle (*null hypothesis significance testing* ou NHST) est l'approche d'inférence statistique la plus largement acceptée et la plus souvent utilisée dans la recherche quantitative en communication. Toutefois, le NHST est très controversé et plusieurs problèmes sérieux de cette approche ont été identifiés. Cet article passe en revue le NHST et la controverse qui l'entoure. Des problèmes couramment reconnus comprennent la sensibilité à la taille de l'échantillon, le fait que la nullité soit généralement littéralement fausse, des taux inacceptables d'erreurs du type II, la mauvaise compréhension et l'abus. Les problèmes associés à la nature conditionnelle du NHST et au défaut de distinguer les hypothèses statistiques des hypothèses réelles sont soulignés. Les solutions et alternatives recommandées sont commentées dans un second article.

Eine kritische Betrachtung des Nullhypothesen-Signifikanztestens in der quantitativen Kommunikationsforschung

Das Nullhypothesen-Signifikanztesten ist die geläufigste und am weitesten akzeptierte Form des statistischen Inferenzschlusses in der quantitativen Kommunikationsforschung. Allerdings ist das Nullhypothesen-Signifikanztesten höchst widersprüchlich und birgt eine Vielzahl gravierender Probleme. Dieser Artikel setzt sich mit dem Nullhypothesen-Signifikanztesten und der damit verbundenen Kontroverse auseinander. Bereits allgemein bekannte Probleme sind die Abhängigkeit von der Stichprobengröße, die Tatsache, dass die Nullhypothese oft buchstäblich falsch ist, außerdem nicht akzeptable Typ II-Fehler-Raten sowie Missverständnisse und Anwendungsfehler. Ergänzend dazu werden Probleme, die auf die Bedingungen des Nullhypothesen-Signifikanztesten zurückzuführen sind sowie das Misslingen zwischen statistischen und substanziellen Hypothesen zu unterscheiden, besprochen. Ein weiterer Aufsatz setzt sich mit möglichen Lösungen und Alternativen auseinander.

Una Evaluación Crítica de la Puesta a Prueba de la Significancia de la Hipótesis Nula en la Investigación de Comunicación Cuantitativa

Timothy R. Levine

Michigan State University

René Weber

University of California, Santa Barbara

Craig Hullett

University of Arizona

Hee Sun Park

Michigan State University

Lisa L. Massi Lindsey

Michigan State University

Resumen

La puesta a prueba de la significancia de la hipótesis nula (NHST) es el enfoque más ampliamente aceptado y usado con frecuencia en la inferencia estadística de la investigación de comunicación cuantitativa. NHST, no obstante, es muy controversial, y varios problemas serios de este enfoque han sido identificados. Este artículo revisa NHST y la controversia que lo rodea. Problemas comúnmente reconocidos son la sensibilidad del tamaño de la muestra, la nulidad es usualmente literalmente falsa, los índices inaceptables de error de tipo II, y el malentendido y abuso. Los problemas asociados con la condición natural de NHST y el fracaso de distinguir las hipótesis estadísticas de las hipótesis sustantivas son enfatizados. Soluciones recomendadas y alternativas son señaladas en un artículo acompañante.

评估定量传播研究中零假设显著性之检测

Timothy Levine
密歇根州立大学
Rene Weber
加州大学 Santa Barbara 分校
Craig Hullett
亚利桑那大学
Hee Sun Park
密歇根州立大学
Lisa Massi Lindsey
密歇根州立大学

零假设显著性检测（NHST）是定量传播研究中使用最频繁、最广为接受的一种统计推理方法。然而，关于 NHST 存在太多的争议；数个严重的问题已被界定。本文对 NHST 及相关争议进行评估。NHST 公认的问题包括易受样本规模影响、零假设常常错误、二类错误比例过高、以及误用及滥用。我们重点探讨与 NHST 条件设置相关的问题以及统计性假设和实质性假设之间的混淆问题。针对这些问题，本文提出解决方案并在另文提供其它选择。

양적 커뮤니케이션 연구에서의 귀무가설 유의성 검증에 관한 비판적 접근

Timothy R. Levine

Rene Weber

Craig Hullett

Hee Sun Park

Lisa L. Massi Lindsey

요약

귀무가설 유의성 검증 (NHST)은 양적 커뮤니케이션 연구에서 통계학적 추론에 대한 접근으로 가장 광범위하게 받아들여지고 자주 사용되어져 왔다. NHST는 그러나 매우 논쟁적인 것이며, 이 접근법에 대한 여러 주요한 문제점들이 제기되어져 왔다. 본 논문은 NHST과 이를 둘러싼 논쟁을 점검하기 위한 것이다. 일반적으로 인지된 문제들은 표본크기에 대한 민감도, 귀무가설이 일반적으로 잘못됐다는 것, 받아들여지지 않은 제2종 오차 비율, 그리고 오해와 오용등에 관한 것이다. NHST의 상황적 본질에 관련된 문제들과 실제가설로부터 통계학적 가정들을 구별하는데 있어서의 잘못들이 강조되었다. 제안적 해결책들과 대안들이 또 다른 연구에서 제기되었다.