Psychophysiology, *33* (1996), 175–183. Cambridge University Press. Printed in the USA.

METHODOLOGY

Effect sizes and *p* values: What should be reported and what should be replicated?

ANTHONY G. GREENWALD,[a] RICHARD GONZALEZ,[a]
RICHARD J. HARRIS,[b] AND DONALD GUTHRIE[c]

[a]Department of Psychology, University of Washington, Seattle, USA
[b]Department of Psychology, University of New Mexico, Albuquerque, USA
[c]Department of Psychiatry, University of California, Los Angeles, USA

Abstract

Despite publication of many well-argued critiques of null hypothesis testing (NHT), behavioral science researchers continue to rely heavily on this set of practices. Although we agree with most critics' catalogs of NHT's flaws, this article also takes the unusual stance of identifying virtues that may explain why NHT continues to be so extensively used. These virtues include providing results in the form of a dichotomous (yes/no) hypothesis evaluation and providing an index (*p* value) that has a justifiable mapping onto confidence in repeatability of a null hypothesis rejection. The most-criticized flaws of NHT can be avoided when the importance of a hypothesis, rather than the *p* value of its test, is used to determine that a finding is worthy of report, and when $p \equiv .05$ is treated as insufficient basis for confidence in the replicability of an isolated non-null finding. Together with many recent critics of NHT, we also urge reporting of important hypothesis tests in enough descriptive detail to permit secondary uses such as meta-analysis.

Descriptors: Replication, Statistical significance, Null hypothesis testing, Methodology

To demonstrate that a natural phenomenon is experimentally demonstrable, we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment [that] will rarely fail to give us a statistically significant result. (Fisher, 1951, p. 14)

Readers of research reports reasonably desire confidence that published results constitute what Fisher called "demonstrable" phenomena—ones that can be confidently reproduced by conscientious and skilled researchers. Such confidence is questionable for the class of isolated findings—ones that are not yet supported by replications. Isolated findings are likely to occur either in initial empirical tests of novel hypotheses or as unpredicted results from exploratory analyses of data.

An isolated finding might be judged as demonstrable (in Fisher's sense) if its empirical support suggests that it can be replicated. Before being able to discuss the relation of empirical evidence to a conclusion about an isolated finding's replicability, it is necessary to clarify what is meant by a finding being replicated or replicable. This, in turn, requires a choice between two approaches to conceiving replication: the choice between considering the meaning of replication in the context of null hypothesis testing (NHT)[1] versus estimation.

Preparation of this report was aided by National Science Foundation grants DBC-9205890, SES-9110572, and SBR-9422242, and by National Institute of Mental Health grant MH-41328.

We thank J. Scott Armstrong, David Bakan, Ronald P. Carver, Jacob Cohen, Alice Eagly, Lynne Edwards, Robert Frick, Gideon Keren, Lester Krueger, Fred Leavitt, Joel R. Levin, Clifford Lunneborg, David Lykken, Paul E. Meehl, Ronald C. Serlin, Bruce Thompson, David Weiss, and this journal's reviewers for comments on an earlier draft. Although these comments helped greatly in sharpening and clarifying arguments, the article's subject matter regrettably affords no uniform consensus, so the final version contains assertions that will be unacceptable to some commenters.

Address reprint requests to: Anthony G. Greenwald at Department of Psychology, Box 351525, University of Washington, Seattle, WA 98195-1525, USA. E-mail: agg@u.washington.edu.

Editor's note. This paper was invited by the Board of Editors of *Psychophysiology* to address standards for reporting data and replication in psychophysiological research. The perspectives that were developed are not specific to psychophysiological data but rather apply to the analysis and interpretation of various forms of empirical data. Although the authors' recommendations regarding statistical methodology might be regarded as controversial, it is our hope that the examples and recommendations offered in this article will focus attention on important issues concerning replicability in the field.

[1] What is called *null hypothesis testing* here is called *null hypothesis significance testing* or simply *significance testing* elsewhere. The label *significance* is deliberately minimized in this article because some of the most objectionable characteristics of NHT stem from misleading uses of that word, which have developed in association with NHT.

In the NHT context, two statistical tests can be said to replicate one another when they support the same conclusion (nonrejection or rejection in a specific direction) with respect to the same null hypothesis.[2] In contrast, in the estimation context two point or interval estimates are said to replicate one another when they meet some criterion of proximity (e.g., overlap of confidence intervals). These two meanings of replication in the NHT and estimation contexts are so divergent that it does not appear to be possible to discuss replication coherently in a general fashion that encompasses both. Although the estimation approach to data analysis and reporting is often preferable to NHT, nevertheless it was easy to choose NHT as the perspective on which to focus in this article. The choice was determined both because of NHT's prevalence in published empirical reports and because the problem of interpreting isolated findings is one that arises especially in the context of NHT.

As will be seen below, there is a large literature of authoritative attacks against NHT. Because there has been relatively little published defense of NHT in response to those attacks, it is remarkable that NHT nevertheless remains virtually uncontested in its dominance both of texts used in basic behavioral science statistics courses and of journals that publish original behavioral science empirical reports. As will be seen, the paradox of this opposition between pundits and practitioners has motivated much of this article, which is organized into three sections. The first section surveys some well-established criticisms of NHT. The second section describes some virtues of NHT that may explain why it remains so popular in the face of sustained heavy criticism. The third section gives recommendations for using NHT in a way that minimizes its flaws while sustaining its virtues.

Criticism of Null Hypothesis Testing (NHT): Three Severe Flaws

Because the Null Hypothesis Is "Quasi-Always" False, Testing It Is Uninformative

In comparing psychology with physics, Meehl (1967; see also 1978) noted that, because of the complexity of influences on measures used in psychological research, psychologists' null hypotheses are almost never exactly true. As Meehl described it, "the point-null hypothesis . . . is [quasi-] always false in biological and social science" (p. 108). Furthermore, empirical tests usually associate the researcher's preferred theoretical prediction with a rejection of the null hypothesis. It follows from the quasi-falsity assumption that, when the researcher's theory is incorrect or irrelevant, there may nevertheless be a good chance that an empirical test will yield a null hypothesis rejection in the predicted direction. That is, assuming quasi-falsity, the null hypothesis should be false in an incorrect or irrelevant theory's predicted direction half of the time, and possibly more, if this direction is plausible from multiple theoretical perspectives. Therefore, the probability that an incorrect or irrelevant theory will be supported by a prediction-confirming null hypothesis rejection approaches 50% (or more) as the test's power increases. Meehl contrasted this analysis of spurious confirmation of theory resulting from increased research power in psychology with a description of statistical practice in physics. In physics, Meehl observed, statistics are used chiefly to estimate theory-specified parameter values from empirical data. In this context of estimation (rather than NHT), increased power yields increased precision and in turn reduces the likelihood of spurious theory confirmations.

In the fashion just described, Meehl targeted NHT as a methodological culprit responsible for spurious theoretical conclusions. Although similar observations had been made previously by (at least) Nunnally (1960), Binder (1963), and Edwards (1965) and although other analysts have argued that the psychology-versus-physics comparison is not so unfavorable to psychology (Hedges, 1987; Serlin & Lapsley, 1993), Meehl's critique nevertheless attracted great attention and continues to be cited as an authoritative statement of the view that NHT is fundamentally flawed because psychological null hypotheses are virtually never true.[3]

NHT Doesn't Tell Researchers What They Want to Know

Several critics have pointed out that the information that researchers should most want to learn from research findings is not, and cannot, be provided by NHT. In particular, NHT does not establish the degree of truth or belief that can be credited to the null hypothesis (Gigerenzer & Murray, 1987; Oakes, 1986; Rozeboom, 1960). NHT does not offer an estimate of the magnitude of a treatment effect or of a relationship between variables (Cohen, 1994; Rosenthal, 1993). And NHT does not provide the probability that a finding can be replicated (Gigerenzer & Murray, 1987; Lykken, 1968; Rosenthal, 1991). Rather, the p value provided by NHT gives only a measure of the probability of results as extreme as (or more extreme than) the obtained data, contingent on the truth of the null hypothesis.[4]

NHT Is Biased Against the Null Hypothesis

A very familiar comment about NHT is that it permits only one conclusive direction of interpretation, that is, against the null hypothesis. The folklore of the field has it that "You can't prove the null hypothesis." In terms of this folklore, what you can do is reject the null hypothesis and conclude that it is wrong, or you can fail to reject the null hypothesis and thereby be left uncertain about its truth. This folkloric assertion is somewhat misleading because it implies, incorrectly, that point hypotheses other than the null hypothesis can be proved. More properly, any point hypothesis (null or otherwise) has equal status in regard to provability. None is provable by the methods of NHT. Generally neglected in this debate is the possibility that, by using Bayesian or estimation methods, any hypothesis (including the null) can gain considerable credibility (see, e.g., Goodman and Royall, 1988; Greenwald, 1975). Nevertheless, there is indeed an asymmetry between the null and alternative hypotheses in NHT, which can be described as follows.

A null hypothesis rejection discredits the null hypothesis and thereby supports alternative hypotheses, but a nonrejection result does not correspondingly support the null hypothesis and

[2] This concept of replication in the NHT context is given a more formal definition below.

[3] Although Meehl's description of the pervasive falsity of psychological point null hypotheses continues to have force, Frick (1995) recently pointed out that there are important cases of psychological research for which it is reasonable to treat the point null hypothesis as true.

[4] In the second section of this article, it is shown that this assertion understates the information provided by p values.

discredit alternatives. There are several reasons for this asymmetry, including (a) the null is typically a point hypothesis, whereas the alternative is typically a set or range of hypotheses; (b) when used in the Neyman–Pearson decision form, probability of Type 2 error (falsely accepting the null) typically exceeds that of Type 1 error (falsely rejecting the null); and (c) perhaps most importantly, nonrejection of the null hypothesis is considered to result plausibly from the researcher's use of invalid research operations.

Greenwald (1975) investigated these and other forms of prejudice against the null hypothesis by conducting a survey of researchers' practices and by constructing a simulation model of the effects of these practices on information transmission characteristics of the research-publication system (see also Oakes, 1986, chapter 1). One of the more important varieties of prejudice against the null hypothesis identified in that review comes about as a consequence of researchers much more often identifying their own theoretical predictions with rejections (rather than with acceptances) of the null hypothesis. The consequence is an ego involvement with rejection of the null hypothesis that often leads researchers to interpret null hypothesis rejections as valid confirmations of their theoretical beliefs while interpreting nonrejections as uninformative and possibly the result of flawed methods. Consistent with this conception of researcher bias, Greenwald's (1975) survey of 75 active social psychological researchers revealed that nonrejections of the null hypothesis are reported as being much less likely (than null hypothesis rejections) to be submitted for publication (1:8 ratio).

Another sense in which NHT is biased against the null hypothesis is captured by the concept of alpha inflation. It is well understood that p values become inaccurate when multiple null hypothesis tests are conducted simultaneously (e.g., Miller, 1981; Selvin & Stuart, 1966). As a minimal example, when an experiment includes two independent tests of the same hypothesis, the probability that at least one of them will achieve an $\alpha \leq .05$ criterion, conditional on truth of the null hypothesis, is not .05 but approximately .10 ($\cong 1 - .95^2$). When multiple null hypothesis tests are conducted, but not all are reported—as when one reports only null hypothesis rejections from multiple tests conducted in a single study, or when one publishes only studies that achieved null hypothesis rejections from a series of studies—the inflation of the reported (nominal) p value has been ignored and the reported value is therefore patently misleading. Despite the known impropriety of this practice, ignoring alpha inflation nevertheless remains a common practice, one that clearly reveals bias against the null hypothesis.

Why Does NHT Remain Popular? Three Reasons

In summary of the preceding section, NHT is multiply misused. In highly powerful studies, null hypothesis rejection is a virtual certainty. Null hypothesis rejections are likely to be overinterpreted as theory confirmations. Null hypothesis tests are misinterpreted as providing information about the plausibility of the null hypothesis. Null hypothesis tests fail to provide the type of information that researchers want most to obtain. In use, NHT is biased toward rejection of the null hypothesis. All of these practices render NHT prone to faulty conclusions that leave readers of research reports, not to mention the NHT-using researchers themselves, misinformed about the empirical status of hypotheses under investigation. Nevertheless, and despite repeated and prominent publication of critiques of NHT that have established these anti-NHT conclusions, behavioral scientists appear not to have been deflected from sustained heavy use of NHT.

One is tempted to conclude that NHT is an addictive affliction of behavioral scientists. Despite repeated admonitions about its harmful effects, researchers not only sustain the behavior of NHT but practice it at the increasing pace afforded by modern computational technology. Perniciously, the harmful effects of using NHT may be felt more by audiences who see products of NHT's practice than by the practicing users themselves, who experience the benefit of an occasional joyous rush toward publication.

The metaphor of the NHT user as an addict may be clever, but it must also be, if not simply inaccurate, at least very incomplete. This conclusion comes partly from observing the absence of impact of NHT-critical literature. This critical literature has been appearing steadily and with relatively little published opposition for more than 30 years. An indicator of the balance of intellectual firepower on the NHT topic can be obtained by counting pro-NHT versus anti-NHT publications in prominent journals. This test indicates a heavy anti-NHT majority. Carver (1993) commented: "During the past 15 years, I have not seen any compelling arguments in defense of statistical significance testing" (p. 287). A prospective test of the present balance of authoritative opinion can be obtained by watching to see whether Cohen's (1994) recent and very prominent broadside against NHT elicits any substantial public defense of NHT.

Despite sustained and consistent expert opposition, NHT persists as a dominant practice among behavioral scientists. In addition, introductory behavioral science statistics courses persist in using texts that are based on NHT. An especially ironic indicator of the lack of impact of anti-NHT critiques occurs in the form of empirical publications by authors of NHT critiques (including one of the present authors), published after the appearance of their critiques. Not invariably, but nevertheless very frequently, these publications reveal extensive use of NHT (e.g., Ambady & Rosenthal, 1993; Eckert, Halmi, Marchi, & Cohen, 1987; Greenwald, Klinger, & Schuh, 1995; Hell, Gigerenzer, Gauggel, Mall, & Müller, 1988; Lykken, McGue, Bouchard, & Tellegen, 1990). Even in a journal for which a prominent editorial "discouraged" the use of NHT (Loftus, 1993, p. 3), examination of the most recent issues indicated use of NHT in *every* empirical article.

Why does NHT not succumb to criticism? For lack of a better answer, it is tempting to credit the persistence of NHT to behavioral scientists' lack of character. Behavioral scientists' unwillingness to renounce the guilty pleasure of obtaining possibly spurious null hypothesis rejections may be like a drinker's unwillingness to renounce the habit of a pre-dinner cocktail.

Now for a change of pace. In contrast to the despairing addiction metaphor that has been developed above, the following paragraphs argue that NHT survives because it provides two advantages to its users: (a) a dichotomous outcome that can be used as the basis for making needed decisions and (b) a measure of confidence in the outcome of a study. The assertion that NHT has these two desirable properties requires reconsideration of two of the anti-NHT criticisms that were described in the first section of this article.

Reason 1: HT Provides a Dichotomous Outcome

Because of widespread adoption of the convention that $p \leq .05$ translates to "statistically significant," NHT can be used to yield

a dichotomous answer (reject or don't reject) to a question about a null hypothesis. This may often be regarded as a useful answer for theoretical questions that are stated in terms of a direction of prediction rather than in terms of the expected value of a parameter. (Estimation is clearly the preferred method for the latter case.) Examination of published empirical reports in most areas of behavioral science reveals that theory-based hypotheses are very often stated in terms of the direction (and without specifying the magnitude) of a difference between treatments or of a relationship between variables. (Such an observation was essential to Meehl's [1967] critique, and the observation continues to be valid.) For these directional predictions, methods that yield answers in continuous form (such as effect size estimates or Bayesian estimates of posterior likelihoods of hypotheses) may be unsatisfying.[5]

An argument for the desirability of dichotomous research outcomes is not to be found in purely statistical reasoning. Rather, this virtue derives from the inescapable need of humans to act – more precisely, to choose between alternative potential paths of action. The following questions hint at the wide variety of situations for which an answer in the form of yes versus no, or go versus no-go, is more valuable than one that, although possibly more informative statistically, is equivocal with respect to action.

- Is this treatment superior to a placebo?
- Does this experimental procedure constitute an effective manipulation of my independent variable (or do I need to look for an alternative)?
- Is the predictive validity of this aptitude test sufficient so that I should add it to the battery used to make my hiring or admissions decision?
- Is the similarity between monozygotic twins on this trait greater than that between dizygotic twins?
- Is there a match between this defendant's DNA and that found at the scene of the crime?

Even though it is not a behavioral scientist's question, this last question, perhaps more than any of the others, justifies the assertions that (a) a point null hypothesis can sometimes be true, and (b) there can be a compelling practical need to convert a statistic's value on some continuous numerical dimension into a yes-no decision. In turn, this observation makes clear that dichotomous outcomes are not an intrinsic property of any statistical test but only of the way in which the test is interpreted by its users. In the case of behavioral sciences, the widespread use of NHT as a basis for decisions depends on the willingness of a research community to treat some agreed-on p value as the boundary for deciding how to act on an empirical result.

[5]An almost reflexive response to this assertion by an estimation advocate could be: "Yes, but the estimation method of computing confidence intervals provides the advantages of estimation while also permitting a dichotomous judgment on the data." To that response, we can only reply "Amen." The recommendations offered in the third section of this article urge uniform reporting of estimation statistics to accompany NHT.

Reason 2: p Value as a Meaningful Common-Language Translation for Test Statistics

Computed p values provide a well-established common translation for a wide variety of statistics used in NHT. The translation to p value is produced in the routine operation of many computerized procedures for calculating t, F, r, β, χ^2, and other familiar statistics. As a consequence, most researchers accumulate far more numerical experience with p values than with the various statistics that are so translated. Furthermore, p values have an informal, intuitive interpretation that is far more readily perceived from the report of a test outcome in p value form than from its report in an unconverted two-dimensional metric that combines degrees of freedom with t or F or r, and so on. In its informal interpretation, the p value is an approximate measure of how surprised we should be by a result when we assume that the theoretical or other basis for predicting it is nonsense. Unlike anything that can be perceived so directly from t, F, or r values (with their associated df), a p value's measure of surprise is simply captured by the number of consecutive zeros to the right of its decimal point. A result with two or more leading zeros in its p value is not easy to dismiss, no matter how misguided its underlying rationale appears to be.

Reason 3: p Value Provides a Measure of Confidence[6] in Replicability of Null Hypothesis Rejections

The published anti-NHT literature is a source of many assertions about the lack of justification for interpreting p values in any of several ways that suggest their providing a continuous *measure* of some interesting property of research outcomes. These assertions are often contestably accurate. In particular, as the investigators cited in the following list and others have pointed out, p values (more properly, their complements) should not be interpreted as (a) the probability that the null hypothesis is false (e.g., Cohen, 1994; Oakes, 1986), (b) the probability that the theory that motivated the empirical test is correct (e.g., Meehl, 1967), (c) the probability that an exact replication will repeat the original finding (Carver, 1978, p. 385; Gigerenzer & Murray, 1987, p. 24), or (d) a measure of any parameter of the population from which subjects were sampled (Bakan, 1966, p. 428).

Although p values are not legitimately interpreted in any of the ways just listed, one can also find authoritative assertions that p values *do* provide continuous measures of some aspect of confidence that can be placed in research findings. The following are some examples.

> [S]urely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p? (Rosnow & Rosenthal, 1989, p. 1277)

> [Editors must make] a judgment with respect to confidence to be placed in the findings – confidence that the results of the experiment would be repeatable under the conditions described. . . . [A]n isolated finding . . . at the .05 level or even the .01 level was frequently judged not sufficiently impressive to warrant archival publication. (Melton, 1962, pp. 553–554).

> It has been argued in this section that the associated probability [i.e., p value] of a significance test is indeed amenable to an inferential interpretation. (Oakes, 1986, p. 29)

[6]This use of *confidence* is its customary meaning found in any dictionary, not to be confused with its usage in the statistical concept of *confidence interval*.

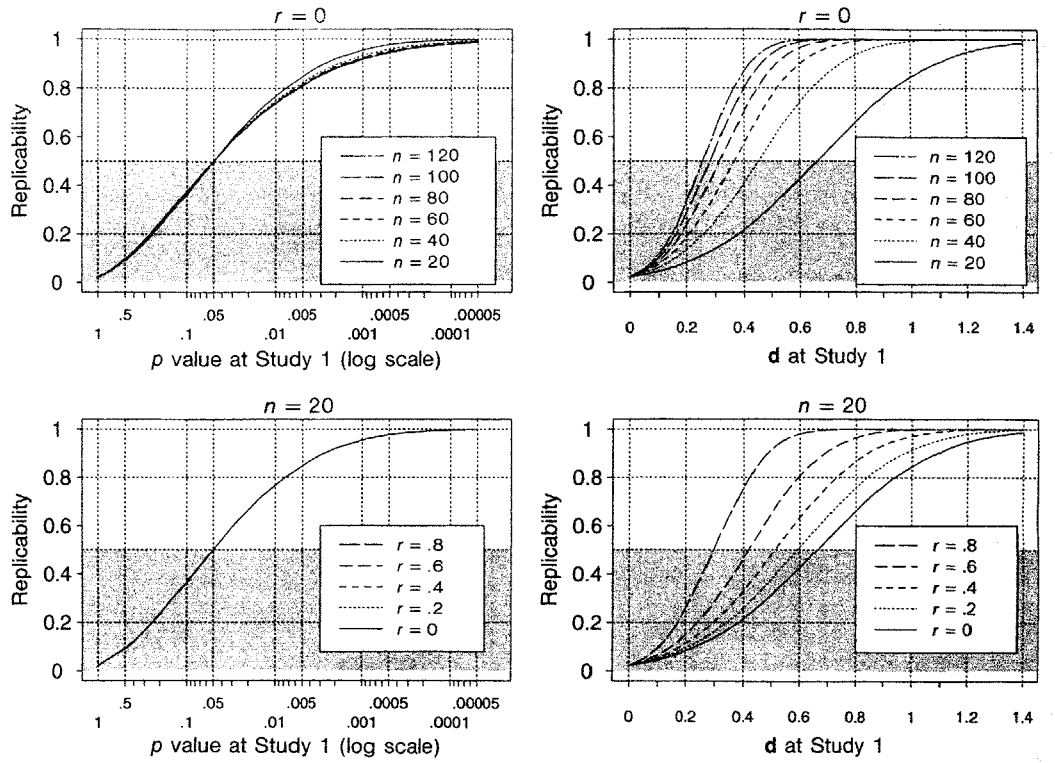

Figure 1. Estimated replicability as a function of p value (log scale) and effect size (**d**), assuming a two-treatment repeated-measures design with selected samples sizes (ns) and selected levels of between-treatment population correlation (r) of the dependent variable. The shaded portion of each plot gives values that are not properly replicabilities because the initial result is not a null hypothesis rejection in these regions. These regions of the plot show the probability of obtaining a null hypothesis rejection (at $\alpha = .05$, two-tailed) from a second study, contingent on the p value of a first study that did not reject the null hypothesis.

In pursuit of the intuitions represented by these remarks, the authors undertook to examine with some care the proposition that p value provides a measure that reflects some aspect of confidence in research outcomes. A demonstration of the validity of that intuition could contribute much to explaining behavioral researchers' attraction to p values. The conclusion that will be reached is that, unlike an effect size (or a confidence interval), a p value resulting from NHT is monotonically related to an estimate of a non-null finding's replicability. In this statement, replicability (which is defined more formally just below) is intended *only* in its NHT sense of repeating the reject-nonreject conclusion and not in its estimation sense of proximity between point or interval estimates.

For purposes of the analysis that leads to results displayed in Figure 1, the NHT sense of replicability is here defined as the estimated probability that an exact replication of an initial null hypothesis rejection will similarly[7] reject the null hypothesis. (To repeat: This definition of replicability is suitable only for the NHT context; it does not capture what is meant by replicability in the estimation context.) As defined in this fashion for the NHT context, replicability can be computed as the power of an exact replication study, which can be approximated with the formula given in Equation 1 (cf. Hays, 1995, p. 329)[8]

$$\text{Power} = 1 - P\left(z \leq \frac{t_{\text{crit}} - t_1}{\sqrt{1 + \dfrac{t_{\text{crit}}^2}{2 \times df}}}\right) \qquad (1)$$

where t_{crit} is the critical t value needed to reject the null hypothesis, df is the degrees of freedom, $P(\,)$ is the probability under the cumulative normal distribution, and t_1 is the observed t value from the initial study.

Equation 1's use of the observed t value from the initial study (in the numerator of the right-hand expression) incorporates an assumption that the effect size to be expected for an exact replication is the effect size observed in the initial study. This

[7]This definition of replicability specifies "*similarly* reject the null hypothesis" to exclude cases in which two studies both reject the null hypothesis but with very different patterns of data (e.g., with opposite directions of difference between means).

[8]Equation 1 is an approximation because it is based on a normal distribution approximation to the noncentral t distribution and because it uses the effect size from the initial study as the estimate of the (unknown) effect size expected for the exact replication. The former introduces very little error, whereas the latter makes it likely that the estimates shown in Figure 1 are too high, as is discussed later in the text. The effect of the normal approximation to the noncentral t was checked by computing selected points in several of the plots in Figure 1; discrepancies between the Equation 1 value and values using the noncentral t were small enough so that plots using the latter would not noticeably deviate from those shown in Figure 1.

generalization from past to future involves a step of inductive reasoning that is (a) well recognized to lack rigorous logical foundation but is (b) nevertheless essential to ordinary scientific activity. Some concerns about the accuracy of this assumption are noted four paragraphs below.

For computations in producing the plots of Figure 1, sample size (and therefore *df*) for the replication study was assumed to be the same as for the initial study. An easily seen direct implication of Equation 1 is that, when the observed *t* for the initial study equals the critical *t* value (and *p* is thus exactly .05), then the numerator of the right-hand term becomes zero, and the power for the replication is .50 ($= P[z \leq 0.0]$). Stated more intuitively, when the expected effect size exactly matches the one needed to achieve $p = .05$, there should be equal chances of exceeding or falling short of that value in an exact replication.

To show the relation of *p* value, sample size, and effect size to replicability, Figure 1 displays estimated replicabilities for a two-treatment repeated-measures design. The *p* values, sample *n*s, and effect size **d**s used as parameters are values assumed to have been observed in the initial study. For the upper left panel, two-tailed *p* values between 1 and .00005 and *n*s of 20, 40, 60, 80, 100, and 120 were converted into t_1 values and *df* ($= n - 1$), and t_{crit} was the tabled critical value for $\alpha = .05$, two-tailed, for the given *df*. The replicabilities on the ordinate are the power values obtained by applying Equation 1. (In these computations, replication success does not mean that the replication matches or exceeds the initial study's *p* value; rather, the criterion for replication of a null hypothesis rejection is uniformly $\alpha = .05$, two-tailed.) The upper right panel was produced similarly, by converting **d**s between 0.0 and 1.4 along with the various *n*s into values of t_1 used in Equation 1. For both upper panels, the correlation between paired observations was assumed to be zero. Note that the two upper plots resemble those that would be obtained from between-subjects designs in which the *n*s given in the legend denote the sample size in each cell. For the lower panels, with *n* fixed at 20, the correlation between paired observations was varied across the levels of $r = 0$, .2, .4, .6, and .8. These increasing correlations leave effect size (**d**) unaltered[9] but increase power substantially, resulting in the varied curves for replicability as a function of *r* and effect size shown in the lower right panel. Because the effect of increasing *r* that increases replication power also decreases the *p* value of the initial study, the five curves of the lower left panel are entirely superimposed, leaving the (incorrect) appearance that four curves have been omitted from the plot.

The difference between the two left and the two right panels of Figure 1 is very striking. The two left panels show that replicability is closely related to the *p* value of an initial study, across variations of sample size and correlation between treatments in a repeated-measures design. By contrast, the two right panels show the lack of any simple relation of effect size to replicability. This contrast provides the basis for observing that the *p* value of a study does provide a basis for estimating confidence in replicability of an isolated finding, when replicability is understood exclusively in its NHT-context interpretation.

In using Figure 1 to estimate the replicability of an isolated finding from its reported *p* value, one must exercise caution. The plotted curves are best regarded as upper bounds of estimated replicability because of one difficult-to-justify assumption that was used in constructing the figure. The difficult-to-justify assumption is that the expected effect size of an exact replication is well estimated by an isolated finding's observed effect size. A more justifiable expectation is that the effect size for an exact replication of an isolated finding is likely to be smaller than the observed effect size. This alternative view can be based on the general principle of expected regression toward the mean and on two more technical lines of argument (the details of which will not be presented here): (a) a maximum likelihood analysis offered by Hedges and Olkin (1985, chapter 14) and (b) a Bayesian analysis whenever the distribution of belief (over hypotheses) prior to the initial study accords greater likelihood to smaller effect sizes than to the observed effect size. Therefore, in using Figure 1 to estimate replicability of an isolated finding, it may be wise to assume that expected replicability is lower than the plotted value.

In conclusion of this section, one of the criticisms of NHT reviewed in the first section of this article was that *p* values do not provide the information that they are sometimes interpreted as providing. It is now apparent that this uninformativeness criticism has been overstated. It has been overstated by writers who have asserted that *p* values cannot be interpreted as providing any information about confidence in the replicability of findings (e.g., Bakan, 1966, p. 429; Carver, 1993, p. 291). In fairness to these writers, they were implicitly or explicitly interpreting replicability in an estimation context, and in that context their assertions *are* accurate. However—and as shown straightforwardly in Figure 1—in the NHT context that continues to prevail both in archival publication and statistics education in the behavioral sciences, *p* value does provide a measure of confidence in replicability of a null hypothesis rejection. Although neither the numerical value of *p* nor its complement (as suggested by Nunnally, 1975) can be interpreted as an estimated probability of replicating a null hypothesis rejection, nevertheless, as can be seen in Figure 1, *p* value does provide a continuous measure that has an orderly and monotonic mapping onto confidence in the replicability of a null hypothesis rejection.

This analysis of NHT should make the typical reader (who is with high probability a user of NHT) feel like the drinker who has just learned that moderate alcohol consumption has desirable effects such as reducing the risk of heart attack (Peele, 1993; Shaper, 1993). Used in moderation, NHT can be of value. The next section gives suggestions for using NHT "in moderation."

Using NHT While Avoiding Its Severe Flaws: Five Recommendations

When used in conjunction with a conventional criterion (such as $\alpha = .05$), NHT's *p* value allows a yes–no appraisal of the accuracy of a directional prediction and can be used to estimate (as in Figure 1) confidence that a similar null hypothesis rejection would result if it were possible for the study to be replicated exactly. For empirical results that are defined in terms of a predicted direction of treatment effect or correlation, these two types of information are directly useful in assessing what Fisher (1951) called *demonstrability*. Because many theory-based hy-

[9] More precisely, effect size is unaltered by variations in *r* when the unit for effect size is the pooled within-treatment standard deviation. This is the effect size measure that is considered most justifiable in meta-analysis, especially when results from between- and within-subjects designs are being combined in estimating a common effect size (Glass, McGaw, & Smith, 1981, p. 117). By contrast, the unit for effect size that is used in computing power is the standard deviation of the difference between treatments.

potheses in behavioral science are formulated only as directional predictions, NHT is justifiably regarded as useful in appraising the results of their empirical tests.

Although useful for evaluating empirical success of directional predictions, NHT nevertheless bears severe flaws for which it has been justly criticized. In most cases, these flaws, which were summarized in the first section of this article, are avoidable accompaniments of NHT that can be minimized by using NHT cautiously and carefully. The remainder of this article attempts to give specific form to this advice.

Recommendation 1: Report p Values With an = Rather Than With a < or a >

This recommendation is first, not because it is the most important but because it follows most directly from the analysis of the preceding section, where it was shown that replicability of a null hypothesis rejection is a continuous, increasing function of the complement of its p value. This attribute of p values makes it preferable to report them numerically rather than either in the more traditional NHT form of relation to an alpha criterion or simply as "significant" or "not significant." That is, it is more informative to describe a result as (say) $p = .003$ rather than as either "$p < .05$" or "statistically significant," and, similarly, it is more useful to describe a result as (say) $p = .07$ rather than as either "$p > .05$" or "not significant." Because the most interpretable characteristic of p values is well captured by counting the number of successive zeros to the right of the decimal point, a useful rule of thumb may be to report p values rounded to the kth decimal place, where k is the position in which the first nonzero digit occurs: For example, $p = .3, .07, .002, .0008, \ldots, 6 \times 10^{-7}$.

Recommendation 2: Treat p ≅ .05 as an Interesting, but Unconvincing, Support for an Isolated NHT Result

From Figure 1 (or Equation 1), it can be seen that $p \cong .05$ translates to an upper bound estimate of 50% chance of obtaining a similar null hypothesis rejection from an exact replication. This 50% figure is well short of any reasonable interpretation of what Fisher (1951) meant by a phenomenon being *demonstrable*. Therefore, when $p \cong .05$ for an isolated finding, an appropriate response on the part of the researcher should be to seek further support by conducting a replication.

This second recommendation should not be treated as an absolute. There will be situations in which an isolated null hypothesis rejection at $p \cong .05$ warrants publication without replication. When the theoretical content of a report is exceptional, evidence for the replicability of the accompanying data may seem a relatively minor consideration in appraising the value of the report. And, when the cost of collecting replication data is very expensive, the information value of reporting precious available evidence may outweigh the more ordinary value of establishing that the observed phenomenon is (in Fisher's sense) demonstrable.

In attempting to replicate an isolated finding for which $p \cong .05$, the researcher should bear in mind that the estimated replicability of 50% shown in Figure 1 is almost certainly an overestimate. It is an overestimate both because (a) as already noted, true effect size will often be lower than the initial study's observed effect size (on which the 50% replicability estimate is based); and (b) the 50% figure assumes an exact replication, which is a practical impossibility (see Caution 1). Furthermore, and as argued by Lykken (1968), the researcher is generally well advised to conduct a replication that varies details of procedures that are irrelevant to the theory or hypothesis under investigation. Success in such a conceptual replication provides greater confidence in the theory underlying the finding than does success in a literal or exact replication.

Recommendation 3: Treat p ≅ .005 as an Indicator of Demonstrability for an Isolated NHT Result

Although other numerical replicability values could be suggested, the most obvious choice for a minimum level that would justify description as "demonstrable" is the 80% figure that Cohen (e.g., 1977, p. 56) advocated as a conventionally acceptable level of statistical power for NHT. Reading the abscissa of Figure 1 to find the p value that corresponds to an estimated replicability of .80 (for an isolated finding) yields $p \cong .005$ for all but the smallest value of sample size in the upper left panel of Figure 1. However, two substantial cautions must be attached to the suggestion that $p \cong .005$ can serve as a p value criterion of demonstrability.

Caution 1. The 80% estimate of replicability that is associated with $p \cong .005$ in Figure 1 applies to the ideal case of an exact replication. An exact replication is a test conducted with additional subjects sampled in the same fashion as those in the initial study and tested under conditions identical to those of the initial study. This can be recognized as a practical impossibility, unless this replication has already been conducted—that is, conducted at the same time as the initial study and with random assignment of subjects either to the initial study or to the replication. Any other replication would be rendered nonidentical at least by the intervening passage of time (cf. the discussion of "history" as a threat to validity by Campbell & Stanley, 1966).

Caution 2. Figure 1's upper bounds become inaccurate to the extent that investigators engage in selective reporting of NHT tests. To the extent that investigators are more likely to report low than high p values or to report null hypothesis rejections rather than nonrejections, the p-value-based estimates of replicability provided by Figure 1 will be too high, possibly much too high.

As the authors have discovered in responses to an earlier draft, this recommendation—its cautions notwithstanding—will be the most controversial assertion in this article. Some will object that $p \cong .005$ is too conservative a criterion for warranting confidence in a result; it poses too great a hurdle for a new result to overcome to earn the approbation afforded by publication. Others, taking just the opposite perspective, will observe that even $p \cong .005$ should be regarded as insufficient to justify confidence in the demonstrability of an isolated finding; no finding should be treated as confidently demonstrated when it has been obtained in only a single study.

Recommendation 4: Report Results for All Important Hypothesis Tests

The selective reporting that prompted Caution 2 is a well-recognized flaw of NHT that was discussed in the first section of this article. There exist several methods (see Miller, 1981) for an upward adjustment of p values to compensate for the alpha inflation that is associated with selective reporting of tests that yield low p values. It is not generally recognized that these adjustments are unnecessary when researchers report all of the null

hypothesis tests that they conduct. For example, consider a researcher who has used NHT in a series of five tests of the same hypothesis, only one of which yielded a null hypothesis rejection. If the report of this research neglects to mention the four nonrejection results, the reader can have no idea that a p value reported at (say) $p = .03$ might be more properly interpreted as $p \cong .15$, a value that is Bonferroni adjusted for alpha inflation (see Kirk, 1995, pp. 119–122) and assumes the five tests to be independent of one another. The misleading lower p value implies (by using Figure 1) a replicability of about 60%, compared with the more appropriate inflation-adjusted estimate of about 30%. If the researcher had instead reported all five tests, readers would not be misled by the single result. More generally, when researchers report all NHT tests, it is possible to use nominal (rather than adjusted) p values without misleading readers—granted, the readers' information processing task becomes more complex.

Unfortunately, the recommendation to report all findings is often difficult to follow. Furthermore, there are several good reasons for reporting results selectively. For example, some portion of the results may have been obtained with flawed procedures, producing results that might deserve to be considered uninformative; editors may request the suppression of indecisive findings; publication space considerations may oblige report of only a subset of findings; and so on. It is therefore frequently left to readers' imaginations to judge the extent to which a report has selectively presented findings. Editors might seek to minimize this problem by asking researchers to report all results for tests of interesting hypotheses, regardless of the fit between result and prediction. However, editors are unlikely to request such full reporting routinely, not only because of the demand that this policy would place on precious publication space but also (more to the point) because whatever their other skills, editors are rarely clairvoyant.

Recommendation 5: Report Enough Data to Permit Secondary Analysis

Even while undertaking to document some virtues of NHT that justify its use with care and caution, the present authors have mentioned a few times that they regard estimation as having generally greater value than NHT. The result of NHT, when reported in minimal form (i.e., as $p \leq \alpha$ or $p > \alpha$), has little use beyond the immediate one of informing a dichotomous reject–nonreject judgment. An accompanying report of either the numerical p value or the numerical value and df of a test statistic such as t, F, or χ^2 (from which p value can be determined) adds to the information value of an NHT report. Reporting of standard deviations of dependent measures further facilitates secondary uses of the reported results (especially in translating findings into effect sizes that can be included in a subsequent meta-analysis). The strong arguments for (a) considering effect sizes in the design and statistical analysis of research and (b) reporting data in detail sufficient to permit secondary use have been made so effectively elsewhere that it suffices here to point readers toward those arguments (e.g., Cohen, 1990; Rosenthal, 1993; Serlin & Lapsley, 1993).

Conclusion

Despite the dominant anti-NHT character of statistical methodological critiques of the past 30 years, NHT retains a tenacious grip on methodological practice in behavior science. This article describes two features of NHT that can explain this paradox. First, NHT allows conversion of test statistics into dichotomous outcomes that can guide decisions in situations that call for practical action. Second, the p value computed in NHT is informative both about the surprise value of a null hypothesis rejection when one disbelieves its theoretical basis and as an indicator of the likelihood that an exact replication would similarly reject the null hypothesis.

This article provides a very limited appreciation of NHT. The unusualness of even this limited appreciation will almost certainly cause the article to be misread as providing unmitigated praise for NHT and the following sentence to be read as less than heartfelt. To the contrary, however, the authors (a) agree with the majority of published criticism of NHT (and have written some of it), (b) greatly dislike statistical analyses that focus more on "statistical significance" than on description of findings and their theoretical or practical importance, (c) regard estimation as a necessary approach for the long-term future of behavioral science, and, as was stated directly in the final section of this article, (d) endorse the reporting of estimation statistics (such as effect sizes, variabilities, and confidence intervals) for all important hypothesis tests.

As Fisher (1951) stated in the passage quoted at the beginning of this article, an NHT result can be regarded as demonstrable when there exists a "reliable method of procedure" that will "rarely fail to give us a statistically significant result." Establishing that a finding is confidently demonstrable in Fisher's (NHT) sense typically depends on conducting and reporting actual replications, even when an isolated finding seems statistically secure. NHT can advise a researcher who observes an isolated null hypothesis rejection as to whether a replication effort is likely to succeed in repeating that result. As this article establishes, the finding that "$p = .05$" falls well short of providing confidence that the isolated finding will "rarely fail" to be observed.

REFERENCES

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*, 431–441.

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

Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 70*, 107–115.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

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.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

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.

Eckert, E. D., Halmi, K. A., Marchi, P., & Cohen, J. (1987). Compar-

ison of bulimic and non-bulimic anorexia nervosa patients during treatment. *Psychological Medicine, 17*, 891–898.
Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin, 63*, 400–402.
Fisher, R. A. (1951). *The design of experiments* (6th ed.). Edinburgh: Oliver & Boyd. (Originally published 1935)
Frick, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition, 23*, 132–138.
Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.
Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
Goodman, S. N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health, 78*, 1568–1574.
Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.
Greenwald, A. G., Klinger, M. R., & Schuh, E. S. (1995). Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General, 124*, 22–42.
Hays, W. L. (1995). *Statistics* (5th ed.). New York: Holt, Rinehart, and Winston.
Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42*, 443–455.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
Hell, W., Gigerenzer, G., Gauggel, S., Mall, M, & Müller, M. (1988). Hindsight bias: An interaction of automatic and motivational factors? *Memory and Cognition, 16*, 533–538.
Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks-Cole.
Loftus, G. R. (1993). Editorial comment. *Memory and Cognition, 21*, 1–3.
Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 151–159.
Lykken, D. T., McGue, M., Bouchard, T. J., & Tellegen, A. (1990). Does contact lead to similarity or similarity to contact? *Behavior Genetics, 20*, 547–561.
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 risks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.
Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64*, 553–557.
Miller, R. G. Jr. (1981). *Simultaneous statistical inference* (2nd ed.). New York: McGraw-Hill.
Nunnally, J. C. (1960). The place of statistics in psychology. *Educational and Psychological Measurement, 20*, 641–650.
Nunnally, J. C. (1975). *Introduction to statistics for psychology and education*. New York: McGraw-Hill.
Oakes, M. (1986). *Statistical inference*. Chichester: Wiley.
Peele, S. (1993). The conflict between public health goals and the temperance mentality. *American Journal of Public Health, 83*, 805–810.
Rosenthal, R. (1991). Cumulating psychology: An appreciation of Donald T. Campbell. *Psychological Science, 2*, 213, 217–221.
Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook of data analysis in the behavioral sciences: Methodological issues* (pp. 519–559). Hillsdale, NJ: Erlbaum.
Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological research. *American Psychologist, 44*, 1276–1284.
Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*, 416–428.
Selvin, H. C., & Stuart, A. (1966). Data-dredging procedures in survey analysis. *American Statistician, 20*, 20–23.
Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook of data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.
Shaper, A. G. (1993). Editorial: Alcohol, the heart, and health. *American Journal of Public Health, 83*, 799–801.

(RECEIVED September 19, 1994; ACCEPTED November 3, 1995)