KLINE Chap. 3

Kline, R.B. (2004). Beyond significance testing. Reforming data analysis methods in behavioral research. Washington, D.C.: APA Books.

3

WHAT'S WRONG WITH STATISTICAL TESTS— AND WHERE WE GO FROM HERE

> Statistics is a subject of amazingly many uses and surprisingly few effective practitioners.
>
> —B. Efron and R. Tabshirani (1993, p. xiv)

This chapter considers problems with null hypothesis significance testing (NHST). The literature in this area is quite large. D. Anderson, Burnham, and W. Thompson (2000) recently found more than 300 articles in different disciplines about the indiscriminate use of NHST, and W. Thompson (2001) lists more than 400 references about this topic. As a consequence, it is possible to cite only a few representative works. General reviews of the controversy about NHST in the social sciences include Borenstein (1998), Nickerson (2000), and Nix and Barnette (1998). Examples of works more critical of NHST include J. Cohen (1994); Gigerenzer (1998a, 1998b); Gliner, Morgan, Leech, and Harmon (2001); and Kruegar (2001), and examples of works that defend NHST include Abelson (1997a, 1997b); Chow (1998a, 1998b); Harris (1997c); and Mulaik, Raju, and Harshman (1997).

After review of the debate about NHST, I argue that the criticisms have sufficient merit to support the minimization or elimination of NHST

The author wishes to thank Richard S. Herrington, as well as the anonymous reviewers, for suggestions in this chapter.

in the behavioral sciences. I offer specific suggestions along these lines. Some concern alternatives that may replace or supplement NHST and thus are directed at researchers. Others concern editorial policies or educational curricula. Few of the recommendations given are original in that many have been made over the years by various authors. However, as a set they deal with issues often considered in separate works. For simplicity, the context for NHST assumed is reject–support (RS) instead of accept–support (AS). The RS context is more common, and many of the arguments can be reframed for the AS context. Exercises for this chapter can be found on this book's Web site.

NHST OUTCOMES ARE INTERPRETED AS SOMETHING THEY ARE NOT

People are by nature good at pattern recognition. We find evidence for this in almost every aspect of human life, whether it is the apparently innate preference of infants for visual stimuli that resemble a human face or the use of language by adults to construct a social reality. There are probably deep evolutionary roots of our ability to find meaning in the world around us. This ability is also at the core of some personality theories. For instance, Rollo May (1975) wrote,

> Creative people . . . do not run away from non-being, but by encountering and wrestling with it, force it to produce being. They knock on silence for answering music; they pursue meaninglessness until they force it to mean. (p. 93)

Our pattern recognition ability is so well-developed that sometimes we see *too much* meaning in otherwise random events. Sagan (1996) described several examples, including one that involved an early satellite photo of a hill in a place called Cydonia on Mars that resembles a human face. Some people took this formation as evidence for a vanished civilization. Later satellite images of the same hill showed pretty clearly that it was carved by natural forces such as wind erosion, but the tendency to see something recognizable in randomness is strong. By virtue of their training or personal dispositions, scientists may be extraordinarily good at pattern recognition, which also makes them subject to the potential error of seeing too much meaning in certain events. This seems to be true about NHST, because many common fallacies about it involve exaggerating what can be inferred from statistical tests. These incorrect inferences may be a source of cognitive misdirection that hinders progress in behavioral research.

Misinterpretations of *p* Values

Next we consider common misunderstandings about the probabilities generated by statistical tests, *p* values. Let us first review their correct interpretation. Recall that statistical tests measure the discrepancy between a sample statistic and the value of the population parameter specified in the null hypothesis, H_0, taking account of sampling error. The empirical test statistic is converted to a probability within the appropriate central test distribution. This probability is the conditional probability of the statistic assuming H_0 is true (see chap. 2, this volume). Other correct interpretations for the specific case $p < .05$ include the following:

1. The odds are less than 1 to 19 of getting a result from a random sample even more extreme than the observed one when H_0 is true.
2. Less than 5% of test statistics are further away from the mean of the sampling distribution under H_0 than the one for the observed result.
3. Assuming H_0 is true and the study is repeated many times, less than 5% of these results will be even more inconsistent with H_0 than the observed result.

That is about it. Other correct definitions may be just variations of those listed. The range of correct interpretations of *p* values is thus actually quite narrow. Let us refer to any correct definition as $p\ (D \mid H_0)$, which emphasizes probabilities from statistical tests as conditional probabilities of the data (D) given the null hypothesis.

Presented next are common misinterpretations for the case $p < .05$. Some of them arise from forgetting that *p* values are conditional probabilities or reversing the two events represented by *p* values, D and H_0. Reasons why each is incorrect are also given below:

Fallacy Number 1

A *p* value is the probability that the result is a result of sampling error; thus, $p < .05$ says that there is less than a 5% likelihood that the result happened by chance. This false belief is the *odds-against-chance fantasy* (Carver, 1978). It is wrong because *p* values are computed under the assumption that sampling error is what causes sample statistics to depart from the null hypothesis. That is, the likelihood of sampling error is *already* taken to be 1.00 when a statistical test is conducted. It is thus illogical to view *p* values as measuring the probability of sampling error. This fantasy together with others listed later may explain the related fallacy that statistical tests

sort results into two categories, those a result of chance (H_0 is not rejected) and others a result of "real" effects (H_0 is rejected). Unfortunately, statistical tests applied in individual studies cannot make this distinction. This is because any decision based on NHST outcomes may be wrong (i.e., a Type I or Type II error).

Fallacy Number 2

A p value is the probability that the null hypothesis is true given the data; thus, $p < .05$ implies $p\,(H_0 \mid D) < .05$. This is the *inverse probability error* (J. Cohen, 1994) or the *Bayesian Id's wishful thinking error* (Gigerenzer, 1993), and it stems from forgetting that p values are conditional probabilities of the data, or $p\,(D \mid H_0)$, and *not* of the null hypothesis, or $p\,(H_0 \mid D)$. The latter is the posterior probability of the null hypothesis in light of the data, and it is probably what researchers would really like to know. A simplified form of Bayes's theorem shows us that $p\,(D \mid H_0)$ from a statistical test and the posterior probability of the null hypothesis are in fact related:

$$p\,(H_0 \mid D) = \frac{p(H_0)\; p(D \mid H_0)}{p(D)} \qquad (3.1)$$

In Equation 3.1, $p\,(H_0)$ is the prior probability that the null hypothesis is true before the data are collected, and $p\,(D)$ is the prior probability of the data irrespective of the truth of the null hypothesis. That is, given the p value from a statistical test along with estimates of $p\,(H_0)$ and $p\,(D)$, we could derive with this equation $p\,(H_0 \mid D)$, the posterior probability of the null hypothesis. Unfortunately, those who use traditional statistical tests do not usually think about prior probabilities. If pressed to specify these values, they may venture a guess, but it may be viewed as subjective. In contrast, a Bayesian approach specifically estimates the posterior probability of the hypothesis, not just the conditional probability of the data under that hypothesis. There are also ways to estimate prior probabilities that are not wholly subjective. Chapter 9 considers the Bayesian approach to hypothesis testing.

Fallacy Number 3

If the null hypothesis is rejected, p is the probability that this decision is wrong; thus, if $p < .05$, there is less than a 5% chance that the decision to reject the null hypothesis is a Type I error. This fallacy is another kind of inverse probability error that Pollard (1993) described as confusing the conditional prior probability of a Type I error, or

$$\alpha = p\,(\text{reject } H_0 \mid H_0)$$

with the conditional posterior probability of a Type I error given that the null hypothesis was rejected, or:

$$p\,(H_0 \mid \text{reject } H_0)$$

Pollard uses Bayes's theorem to show it is not generally possible to estimate $p\,(H_0 \mid \text{reject } H_0)$ from α. On a more intuitive level, the decision to reject the null hypothesis in an individual study is either correct or incorrect, so no probability is associated with it. Only with sufficient replication could we discern whether a specific decision to reject H_0 was correct.

Fallacy Number 4

The complement of p, $1 - p$, is the probability that the alternative hypothesis is true given the data, or $p\,(H_1 \mid D)$. Thus, $p < .05$ says that the likelihood that H_1 is true is greater than 95%. This erroneous idea is the *validity fallacy* (Mulaik et al., 1997) or the *valid research hypothesis fantasy* (Carver, 1978). The complement of p is a probability, but it is just the likelihood of getting a result even *less* extreme under H_0 than the one actually found. Accordingly, complements of p have nothing directly to do with the posterior probability of H_1.

Fallacy Number 5

The complement of p is the probability that a result will replicate under constant conditions; thus, $p < .05$ says that the chance of replication exceeds 95%. This is the *replicability* or *repeatability fallacy* (Carver, 1978). Another variation for $p < .05$ is that a replication has a 95% probability of yielding a statistically significant result, presumably in the same direction as in the original study. If this fallacy were true, knowing the probability of finding the same result in future replications would be very useful. Alas, a p value is just the probability of a particular result under a specific hypothesis. As noted by Carver, replication is a matter of experimental design and whether an effect actually exists in the population. It is thus an empirical question for future studies and not one directly addressed by statistical tests in a single study.

Readers should note, however, that there is a sense in which p values concern replication. Greenwald, Gonzalez, Harris, and Guthrie (1996) made the point that p values in an original study are *monotonically* related to the statistical power of replications. A monotonic relation is typically ordinal or nonlinear; thus, there is not a uniform correspondence between p values and the probabilities of null hypothesis rejections in replications. Specifically, without special graphs like ones presented by Greenwald et al., one cannot directly convert a p value to the likelihood of repeating a null

hypothesis rejection. This is a subtle point. It is probably best to keep in mind that p values have little to do with replication in the usual scientific sense.

Mistaken Conclusions After Making a Decision About the Null Hypothesis

There are also many false conclusions that may be reached after deciding to reject or fail to reject H_0 based on p values. Most require little explanation about why they are wrong:

Fallacy Number 1

A p value is a numerical index of the magnitude of an effect; thus, low p values indicate large effects. This misconception could be called the *magnitude fallacy*. Smaller p values indicate lower conditional probabilities of the data, given the required assumption that the null hypothesis exactly describes the population (J. Cohen, 1994), but that is about all that can be said without other kinds of analyses such as effect size estimation. This is because statistical tests and their p values measure sample size *and* effect size (e.g., Table 2.2), so an effect of trivial magnitude needs only a large enough sample to be statistically significant. If the sample size is actually large, low p values just confirm a large sample, which is a tautology (B. Thompson, 1992). Now, results that are truly of large magnitude may also have low p values—it is just that one cannot tell much by looking at p values alone.

Fallacy Number 2

Rejection of the null hypothesis confirms the alternative hypothesis and the research hypothesis behind it. This *meaningfulness fallacy* actually reflects two conceptual errors. First, the decision to reject H_0 in a single study does not imply that H_1 is "proven." Second, even if the *statistical* hypothesis H_1 is correct, it does not mean that the *substantive* hypothesis behind it is also correct. For example, Arbuthnot (1710) studied the birth records for London for 82 consecutive years (1629–1710). More boys than girls were born every single year during this time. For example, in 1629 there were 5,218 registered births of boys compared with 4,683 births of girls. Based on all these data, Arbuthnot rejected the hypothesis that equal proportions of babies are boys versus girls. In modern terms, he rejected the non-nil hypothesis H_0: $\pi = .50$, where π is the population proportion of boys in favor of the directional alternative hypothesis H_1: $\pi > .50$. However, Arbuthnot's substantive hypothesis was that because of divine providence, more boys are born to compensate for higher numbers of male deaths in wars, accidents, and the like so that, in the end, "every Male may have a Female of the same Country and suitable Age" (1710, p. 188). Arbuthnot was correct about the *statistical* hypothesis H_1, but his substantive hypothesis, although colorful, does not correspond to the actual underlying cause of unequal numbers of newborn boys versus girls: Sperm with Y chromosomes swim faster than those with X chromosomes and arrive in greater numbers to fertilize the egg.

The distinction between statistical and substantive hypotheses is crucial. They differ not only in their levels of abstraction (statistical: lowest; scientific: highest), but also have different implications following rejection of H_0. If H_0 and H_1 reflect only statistical hypotheses, there is little to do after rejecting H_0 except replication. However, if H_1 stands for a scientific hypothesis, the work just begins after H_0 is rejected. Part of the work involves pitting the research hypothesis against other substantive hypotheses also compatible with the statistical hypothesis H_1. If these other hypotheses cannot be ruled out, the researcher's confidence in the original hypothesis must be tempered. It may also be necessary to conduct additional studies that attempt to falsify equivalent models. This is the strategy of *strong inference* (Platt, 1964).

Fallacy Number 3

Failure to reject a nil hypothesis means that the population effect size is zero. This is not a valid inference for a few reasons. One is the basic tenet of science that absence of evidence is not evidence of absence. Also, the decision to fail to reject a nil hypothesis may be a Type II error. For example, there may be a real effect, but the study lacked sufficient power to detect it. Given the relatively low overall power of behavioral research, this is probably not an infrequent event. Poor research design or use of flawed measures can also lead to Type II errors.

Fallacy Number 4

Failure to reject the nil hypothesis H_0: $\mu_1 = \mu_2$ means that the two populations are equivalent. Suppose that an established treatment known to be effective is compared with a new treatment that costs less. It is *incorrect* to conclude that the two treatments are equivalent if the nil hypothesis H_0: $\mu_1 = \mu_2$ is not rejected. The inference of equivalence would be just as incorrect if this example concerned reliability coefficients or proportions in two groups that were not statistically different (Abelson, 1997a; B. Thompson, 2003). To rephrase the tenet cited earlier, the absence of evidence for differences is not evidence for equivalence. Proper methods for equivalence testing are described later.

Fallacy Number 5

Rejecting the null hypothesis confirms the quality of the experimental design. Poor study design can create artifactual effects that lead to incorrect rejection of H_0. Also, plain old sampling error can lead to Type I errors even in well-controlled studies.

Fallacy Number 6

If the null hypothesis is not rejected, the study is a failure. This misconception is the mirror image of the preceding one. Although improper methods or low power can cause Type II errors, failure to reject H_0 can also be the product of good science. For example, some claims based on initial studies are incorrect, which means that replication will lead to negative results. Readers may recall an announcement a few years ago by researchers who claimed to have produced cold fusion (a low energy nuclear reaction) with a relatively simple laboratory apparatus. Other scientists were unable to replicate the phenomenon, and the eventual conclusion was that the claim was premature (Taubes, 1993).

Fallacy Number 7

Rejection of H_0 means that the underlying causal mechanism is identified. This misinterpretation is related to the ones discussed to this point. It should be obvious by now that a single H_0 rejection does not prove a presumed causal effect represented by the statistical hypothesis H_1.

Fallacy Number 8

The failure to replicate is the failure to make the same decision about H_0 across studies. P. Dixon and O'Reilly (1999) refer to this idea as the *reification fallacy*. Under this sophism, a result is considered not replicated if H_0 is rejected in the first study but not in the second study. However, this view ignores sample size, effect size, and the direction of the effect across the two studies. Suppose a group mean difference is found in an initial study and a nil hypothesis is rejected. The exact same group mean difference is found in a replication study, but H_0 is not rejected because of a smaller sample size. We actually have positive evidence for replication even though different decisions about H_0 were made across the two studies.

Widespread Nature of Misinterpretations

There is evidence that many of the false beliefs just described are common even among professional researchers and educators. For instance, Oakes (1986) asked 70 academic psychologists to state their usually adopted

TABLE 3.1
Usually Adopted Interpretations of $p < .01$ by 70 Academic Psychologists

Statement	*f*	%
1. The null hypothesis is absolutely disproved.	1	1.4
2. The probability of the null hypothesis has been found.	32	45.7
3. The experimental hypothesis is absolutely proved.	2	2.9
4. The probability of the experimental hypothesis can be deduced.	30	42.9
5. The probability that the decision taken is wrong is known.	48	68.6
6. A replication has a .99 probability of being significant.	24	34.3
7. The probability of the data given the null hypothesis is known.	8	11.3

Note. From *Statistical Inference* (p. 81), by M. Oakes, 1986, New York: Wiley. Copyright 1986 by John Wiley and Sons. Reprinted with permission.

interpretations of $p < .01$. The respondents could offer more than one interpretation. Of the seven statements listed in Table 3.1, only the last is correct, but just 8 of 70 participants (11%) reported it. Almost 50% endorsed statements 2 and 4 in the table that p values indicate the conditional probability of H_0 (inverse probability error) or H_1 (valid research hypothesis fallacy), respectively. The majority of the respondents said in error that p values are posterior probabilities of Type I error, and about one third said that the complements of p values indicate the likelihood of replication (repeatability fallacy).

Lest one think that Oakes's results are specific to an unrepresentative group of NHST-challenged academic psychologists, results of other surveys of professionals or near-professionals in the social sciences indicate similar, apparently widespread misunderstandings (e.g., Mittag & B. Thompson, 2000; Nelson, R. Rosenthal, & Rosnow, 1986). Tversky and Kahneman (1971) described a kind of cognitive distortion among psychologists they called the *belief in the law of small numbers*. This belief holds that (a) even small samples are typically representative of their respective populations, and (b) statistically significant results are likely to be found in replication samples half the size of the original. The belief in the law of small numbers is probably just as widespread in other social science disciplines as in psychology.

One also need not look very hard in published sources to find errors similar to those in Table 3.1. J. Cohen (1994) listed several distinguished authors who have made such mistakes in print, including himself. This book probably contains similar kinds of errors. Dar, Serlin, and Omer (1994) noted that several prominent psychotherapy researchers who published in some of the best peer-reviewed journals in this area made similar mistakes over a period of three decades. At first glance, this situation seems puzzling. After all, many academicians and researchers have spent hundreds of hours studying or teaching NHST in statistics courses at both the undergraduate

and graduate levels. Why does this rather large investment of educational resources and effort not have more apparent success?

Two factors warrant comment. The first is that NHST is not the most transparent of inference systems. Pollard and others noted that it is difficult to explain the logic of NHST and dispel confusion about it. Some of the language of NHST is very specific and unnatural. For example, the word *significant* implies in natural language that something is important, noteworthy, or meaningful, but not in NHST. There may also be inherent contradictions in the hybrid of the Fisher and Neyman-Pearson models on which contemporary NHST is based (P. Dixon & O'Reilly, 1999; Gigerenzer, 1993). Another problem is a general human weakness in reasoning with conditional probabilities, especially ones best viewed from a relative frequency perspective (e.g., J. Anderson, 1998).

NHST DOES NOT TELL US WHAT WE REALLY WANT TO KNOW

Many of the fallacies about NHST outcomes reviewed concern things that researchers really want to know, including the probability that H_0 or H_1 is true, the likelihood of replication, and the chance that the decision taken to reject H_0 is wrong, all given the data. Using R to stand for replication, this wish list could be summarized as:

$$p\,(H_0 \mid D),\ p\,(H_1 \mid D),\ p\,(R \mid D),\ \text{and}\ p\,(H_0 \mid \text{Reject } H_0)$$

Unfortunately, statistical tests tells us only $p\,(D \mid H_0)$. As noted by J. Cohen (1994), however, there is no statistical technique applied in individual studies that can fulfill this wish list. (A Bayesian approach to hypothesis testing is an exception; see chap. 9, this volume.) However, there is a method that *can* tell us what we really want to know, but it is not a statistical technique; rather, it is replication, which is not only the best way to deal with sampling error, but replication is also a gold standard in science (see chap. 2, this volume). This idea is elaborated next and again in chapter 8.

NIL HYPOTHESES ARE USUALLY FALSE

Nil hypotheses are the most common type tested in the social sciences. However, it is very unlikely that the value of *any* population parameter is exactly zero, especially if zero implies the complete absence of an effect, association, or difference (e.g., Kirk, 1996). For example, the population correlation (ρ) between any two variables we would care to name is probably

not zero. It is more realistic to assume nonzero population associations or differences (see chap. 2, this volume). Meehl (1990) referred to these nonzero effects as a "crud factor" because, at some level, everything is related to everything else; Lykken's (1968) term *ambient correlational noise* means basically the same thing. Although exact values of the crud factor are unknown, correlations may depart even further from zero for variables assessed with the same measurement method. Correlations that result in common method variance may be as high as .20 to .30 in absolute value.

If nil hypotheses are rarely true, rejecting them requires only sufficiently large samples. Accordingly, (a) the effective rate of Type I error in many studies may be essentially zero, and (b) the only kind of decision error is Type II. Given that power is only about .50 on average, the typical probability of a Type II error is also about .50. F. Schmidt (1992, 1996) made the related point that methods to control experimentwise Type I error, such as the Bonferroni correction, may reduce power to levels even lower than .50. It should be said that, as point hypotheses, non-nil hypotheses are no more likely to be true than nil hypotheses. Suppose that a non-nil hypothesis is H_0: $\rho = .30$. The true value of the population correlation may be just as unlikely to be exactly .30 as zero. However, non-nil hypotheses offer a more realistic standard against which to evaluate sample results, when it is practical to actually test them.

Perhaps most p values reported in the research literature are associated with null hypotheses that are not plausible. For example, D. Anderson et al. (2000) reviewed the null hypotheses tested in several hundred empirical studies published from 1978 to 1998 in two prominent environmental sciences journals. They found many biologically implausible null hypotheses that specified things such as equal survival probabilities for juvenile and adult members of a species or that growth rates did not differ across species, among other assumptions known to be false before the data were collected. I am unaware of a similar survey of null hypotheses in the social sciences, but it would be surprising if the results would be appreciably different.

SAMPLING DISTRIBUTIONS OF TEST STATISTICS ASSUME RANDOM SAMPLING

Lunneborg (2001) described this issue as a mismatch between statistical analysis and design. The p values for test statistics are estimated in sampling distributions that assume random sampling from known populations. These are the same distributions in which standard errors for traditional confidence intervals are estimated. Random sampling is a crucial part of the *population inference model*, which concerns the external validity of sample results. However, most samples in the social sciences are not randomly selected—

they are samples of convenience. In experimental studies, it is the *randomization model*, which involves the random assignment of locally available cases to different conditions, that is much more common than the population inference model. Reichardt and Gollob (1999) suggested that results of standard statistical tests yield standard errors that are too conservative (too large) when randomized cases are from convenience samples. They described a modified t test that assumes the population size equals total number of cases, $N = n_1 + n_2$. Lunneborg (2001) described the use of bootstrapping to construct empirical sampling distributions for randomization studies based on convenience samples. Bootstrapping is described in chapter 9.

Bakan (1966) argued that the ideal application of NHST is manufacturing, not the social sciences. Essentially any manufacturing process is susceptible to random error. If this error becomes too great, such as when pistons are made too big relative to the cylinders in car engines, products fail. In this context, the null hypothesis represents a product specification that is reasonable to assume is true, samples can be randomly selected, and exact deviations of sample statistics from the specification can be accurately measured. It may also be possible in this context to precisely estimate the costs of certain decision errors. All of these conditions rarely hold in behavioral research. As the saying goes, one needs the right tool for the right job. Perhaps NHST is just the wrong tool in many behavioral studies.

STATISTICAL ASSUMPTIONS OF NHST METHODS ARE INFREQUENTLY VERIFIED

Statistical tests usually make certain distributional assumptions. Some are more critical than others, such as the sphericity requirement of the dependent samples F test. If critical assumptions are violated, p values may be wrong. Unfortunately, it seems that too many researchers do not provide evidence about whether distributional assumptions are met. H. Keselman et al. (1998) reviewed more than 400 analyses in studies published from 1994 to 1995 in major education research journals, and they found relatively few articles that verified assumptions of statistical tests. Max and Onghena (1999) found a similar neglect of statistical issues across 116 articles in speech, language, and hearing research journals. These surveys reflect an apparently substantial gap between NHST as described in the statistical literature and its use in practice. Results of more quantitative reviews also suggest that there may be relatively few instances in practice when widely used methods such as the standard F test give accurate results because of violations of assumptions (e.g., Lix, J. Keselman, & H. Keselman, 1996).

NHST BIASES THE RESEARCH LITERATURE

There is a sense that journal editors are not interested in publishing studies without H_0 rejections. This perception is supported by (a) comments by past editors of respected journals about favoring studies with H_0 rejections (e.g., Melton, 1962); (b) survey results that show that behavioral researchers are unlikely to submit studies without H_0 rejections for publication (e.g., Greenwald, 1975); and (c) the more causal observation that the large majority of published studies contain H_0 rejections. The apparent bias for studies with statistically significant results presents the difficulties enumerated and discussed next:

1. *The actual rate of Type I error in published studies may be much higher than indicated by* α. Suppose that a treatment is no more effective than control (the nil hypothesis is true) and 100 different studies of the treatment are each conducted at $\alpha = .05$. Of the 100 t tests of the treatment versus control mean contrasts, a total of five are expected to be statistically significant. Suppose these five studies are published, but authors of the other 95 decide not to submit their studies or do so but without success. The actual rate of Type I error among the five published studies is 100%, not 5%. Also, the only studies that got it right—the 95 where H_0 was not rejected—were never published. Clark (1976) made a similar point: Because researchers find it difficult to get their failures to replicate published, Type I errors, once made, are difficult to correct.
2. *The reluctance to submit or publish studies with no statistically significant results leads to a "file drawer problem."* This term is from R. Rosenthal (1979), and it refers to studies not submitted for publication or presented in another forum, such as conferences. It is thought that many file drawer studies contain no H_0 rejections. If an effect is actually zero, results of such studies are more scientifically valid than published studies that reject nil hypotheses.
3. *Published studies overestimate population effect sizes.* Without large samples to study small- or medium-sized effects, it may be difficult to get statistically significant results because of low power. When H_0 is rejected, it tends to happen in samples where the observed effect size is larger than the population effect size. If only studies with H_0 rejections are published, the magnitude of the population effect size winds up being overestimated. An example illustrates this point. Table 3.2

TABLE 3.2
Results of Six Hypothetical Replications

Study	$M_1 - M_2$	s_1^2	s_2^2	t (38)	Reject nil hypothesis?	95% CI
1	2.50	17.50	16.50	1.91	No	−.53–6.53
2	4.00	16.00	18.00	3.07	Yes	1.36–6.64
3	2.50	14.00	17.25	2.00	No	−.03–5.03
4	4.50	13.00	16.00	3.74	Yes	2.06–6.94
5	5.00	12.50	16.50	4.15	Yes	.56–7.44
6	2.50	15.00	17.00	1.98	No	−.06–5.06
Average:	3.58				Range of overlap:	2.06–5.03

Note. For all replications, $n = 20$, $\alpha = .05$, and H_1 is nondirectional. CI = confidence interval.

summarizes the results of six different hypothetical studies where two of the same conditions are compared on the same outcome variable. Note that results of the independent samples t test leads to rejection of a nil hypothesis in three studies (50%), but not in the rest. More informative than the number of H_0 rejections is the average value of $M_1 - M_2$ across all six studies, 3.58. This result may be a better estimate of $\mu_1 - \mu_2$ than the mean difference in any individual study. Now suppose that results from the three studies with H_0 rejections in the table (studies 2, 4, and 5) are the only ones published. The average value of $M_1 - M_2$ for these three studies is 4.22, which is greater than the average based on all six studies.

NHST MAKES THE RESEARCH LITERATURE DIFFICULT TO INTERPRET

If there is a real effect but power is only .50, about half the studies will show positive results (H_0 rejected) and the rest negative results (H_0 not rejected). If somehow all studies are published, the box score of positive and negative results will be roughly equal. From this perspective, it would appear that the research literature is inconclusive (e.g., Table 3.2). Because power is generally about .50 in the social sciences, it is not surprising that only about half of the studies in some areas yield positive results (F. Schmidt, 1996). This is especially true in "soft" behavioral research where theories are neither convincingly supported or discredited but simply fade away as researchers lose interest (Meehl, 1990). Part of the problem comes from interpreting the failure to reject a nil hypothesis as implying a zero population

effect size. Such misinterpretation may also lead to the discarding of treatments that produce real benefits.

There may be other negative consequences of using NHST outcomes to sort studies by whether their results are statistically significant. I have heard many psychology students say, "Research never proves anything." These same students have probably recognized that "the three most commonly seen terms in the [soft social science] literature are 'tentative,' 'preliminary,' and 'suggest.' As a default, 'more research is needed' " (Kmetz, 2000, p. 60). It is not only a few students who are skeptical of the value of research. Clinical psychology practitioners surveyed by Beutler, R. Williams, Wakefield, and Entwistle (1995) indicated that the clinical research literature was not relevant for their work. Similar concerns about research relevance have been expressed in education (D. W. Miller, 1999). These unenthusiastic views of research are the antithesis of the attitudes that academic programs try to foster.

NHST DISCOURAGES REPLICATION

Although I am unaware of data that supports this speculation, a survey would probably find just as many behavioral researchers as their natural science colleagues who would endorse replication as a critical activity. Nevertheless, there is a sense that replication is given short shrift in the social sciences compared to the natural sciences. There is also evidence that supports this concern. Kmetz (1998) used an electronic database to survey about 13,000 articles in the area of organizational science and about 28,000 works in economics. The rates of studies specifically described as replications in each area were .32% and .18%, respectively. Comparably low rates of nominal replications have also been observed in psychology and education journals (e.g., Shaver & Norton, 1980).

The extensive use of NHST in the social sciences and resulting cognitive misdirection may be part of the problem. For example, if one believes that $p < .01$ implies that the result is likely to be repeated more than 99 times out of 100, why bother to replicate? A related cognitive error is the belief that statistically significant findings should be replicated, but not ones for which H_0 was not rejected (F. Schmidt & Hunter, 1997). That NHST makes research literatures look inconclusive when power is low may also work against sustained interest in research topics.

Perhaps replication in the behavioral sciences would be more highly valued if confidence intervals were reported more often. Then readers of empirical articles would be able to see the low precision with which many studies are conducted. That is, the widths of confidence intervals for behavioral data are often, to quote J. Cohen (1994, p. 1002), "so embarrassingly

large!" Relatively wide confidence intervals indicate that the study contains only limited information, a fact that is concealed when only results of statistical tests are reported (F. Schmidt & Hunter, 1997). This reality is acknowledged by the aspect of meta-analytic thinking that does not overemphasize outcomes of statistical tests in individual studies (see chap. 1, this volume).

NHST OVERLY AUTOMATES THE REASONING PROCESS

Social science researchers and students alike seem to universally understand the importance of precise operationalization. The method of NHST offers many of the same apparent advantages in the realm of inference: It is a detailed, step-by-step procedure that spells out the ground rules for hypothesis testing (see chap. 2, this volume). It is also a public method in that its basic rules and areas for researcher discretion are known to all. One of the appeals of NHST is that it automates much of the decision-making process. It may also address a collective need in the social sciences to appear as objective as the natural sciences. However, some critics claim that too much of our decision making has been so automated. Some of the potential costs of letting statistical tests do our thinking for us are summarized next.

1. *Use of NHST encourages dichotomous thinking.* The ultimate outcome of a statistical test is dichotomous: H_0 is either rejected or not rejected. This property may encourage dichotomous thinking in its users, and nowhere is this more evident than for p values. If $\alpha = .05$, for instance, some researchers see a result where $p = .06$ as qualitatively different than one where $p = .04$. These two results lead to different decisions about H_0, but their p values describe essentially the same likelihood of the data (Rosnow & R. Rosenthal, 1989). More direct evidence of dichotomous thinking was described by Nelson et al. (1986), who asked researchers to rate their confidence in results as a function of p values. They found a relatively sharp decline in rated confidence when p values were just above .05 and another decline when p values were just above .10. These changes in confidence are out of proportion to changes in continuous p values.

   That NHST encourages dichotomous thinking may also contribute to the peculiar practice to describe results where p is just above the level of α as "trends" or "approaching significance." These findings are also typically interpreted along with statistically significant ones. However, results with p values just lower than α, such as $p = .04$ when $\alpha = .05$, are almost never described as "approaching nonsignificance" and subsequently discounted. There is a related tendency to attribute the failure to reject H_0 to poor experimental design rather than to the invalidity of the substantive hypothesis behind H_1 (Cartwright, 1973).
2. *Use of NHST diverts attention away from the data and the measurement process.* If researchers become too preoccupied with H_0 rejections, they may lose sight of other, more important aspects of their data, such as whether the variables are properly defined and measured. There is a related misconception that reliability is an attribute of tests rather than of the scores for a particular population of examinees (B. Thompson, 2003). This misconception may discourage researchers from reporting the reliabilities of their own data. Interpretation of effect size estimates also requires an assessment of the reliability of the scores (Wilkinson & the Task Force on Statistical Inference [TFSI], 1999).
3. *The large investment of time to learn NHST limits exposure to other methods.* There is a large body of statistical methods other than NHST that can deal with a wide range of hypotheses and data, but social science students generally hear little about them, even in graduate school. The almost exclusive devotion of formal training in statistics to NHST leaves little time for learning about alternatives. Those who become professional researchers must typically learn about these methods on their own or in workshops.
4. *The method of NHST may facilitate research about fad topics that clutter the literature but have little scientific value.* Meehl's (1990) observations on soft psychology research topics with short shelf lives were mentioned earlier. The automatic nature of NHST has been blamed by some authors as a contributing factor: With very little thought about a broader theoretical rationale, one can collect data from a sample of convenience and apply statistical tests. Even if the numbers are random, some of the results are expected to be statistically significant. The objective appearance and mechanical application of NHST may lend an air of credibility to studies with otherwise weak conceptual foundations.

NHST IS NOT AS OBJECTIVE AS IT SEEMS

The level of α and the forms of H_0 (nil versus non-nil) and the alternative hypothesis (directional versus nondirectional) should be specified before the data are collected. This does not always happen in practice. Under a strict view, this is paramount to cheating. Even under a less demanding standard, the ability to change the rules to enhance the outcome makes the whole process seem more subjective than objective. Selective reporting of results, such as only those where H_0 was rejected, presents a similar problem.

MANY NHST METHODS ARE MORE CONCERNED WITH GROUPS THAN INDIVIDUALS

Statistical tests that analyze means, such as t and F, are concerned with group statistics. They provide little information about individuals within these groups. Indeed, within-groups variances contribute to the error terms of both t and F. However, there are times when it is crucial to understand the nature of individual differences within groups. For instance, it can happen that the group mean difference is statistically significant, but there is substantial overlap of the two frequency distributions. This suggests that the difference at the group level does not trickle down to the case level. Some methods of effect size estimation introduced in chapter 4 analyze differences at the case level.

NHST AND SCHOOLS OF PROBABILITY

In the fields of mathematics, statistics, and philosophy of science, there are several different schools of thought about probabilities, including classical, frequentist, and subjective, among others. There are also deep and long-standing divisions between these schools about the exact meaning of probabilities and their proper interpretation. These debates are complex and highly nuanced, and whole books have been written on the subject (e.g., Hogben, 1957). For these reasons, these debates cannot be summarized in this section. However, readers should know that NHST is associated with only some of these schools of thought about probability; specifically, ones that view probabilities as relative frequencies of repeatable events that can be empirically observed or approximated with theoretical sampling distributions. The method of NHST also uses little previous knowledge other than to assume that H_0 is true. But in no way does NHST represent a consensual view of probability either within or outside the social sciences.

CONTINUED USE OF NHST IS A RESULT OF INERTIA

Several critics have described the continued use of NHST as an empty, ritualized practice, one carried out with little reflection. Education in social science statistics that fails to inform about alternatives may encourage the belief that there is no other way to test hypotheses (F. Schmidt & Hunter, 1997). This belief is unfounded. It is also worth noting that some of the most influential work in psychology, including that of Piaget, Pavlov, and Skinner, was conducted without rejecting null hypotheses (Gigerenzer, 1993). The natural sciences have thrived despite relatively little use of statistical tests.

Others note the general difficulty of changing established methods in science. A familiar, well-entrenched method is like a paradigm, and changing paradigms is not quick or easy (Kuhn, 1996). Such change sometimes awaits the passing of an older generation of scholars and its replacement with younger colleagues who are not as set in their ways. Recall that the adoption of NHST as the standard for hypothesis testing in psychology took about 20 years (see chap. 1, this volume).

IS THERE ANYTHING RIGHT WITH NHST?

The litany of criticisms of NHST reviewed in this chapter raise the question of whether there is anything right about NHST. However, NHST is not without its defenders. Some positive aspects of NHST are enumerated and discussed next.

1. *If NHST does nothing else, it addresses sampling error.* Sampling error is one of the core problems of behavioral research. For all the limitations of p values, they are at least derived taking account of sampling error. Accordingly, some behavioral researchers see NHST as addressing an important need and thus may be less like passive followers of tradition than supposed by critics. Any proposed alternative to NHST must deal with the problem of sampling error lest it be seen as irrelevant to the needs of these researchers. Critics of NHST rightly point out that confidence intervals convey more information about sampling error than test statistics and p values. They also suggest that excessive preoccupation with statistical tests is one reason why confidence intervals are not reported more often. However, confidence intervals are subject to some of the same kinds of inference errors as NHST. Abelson (1997a) made this point in a lighthearted way by describing the "law

of diffusion of idiocy," which says that every foolish practice of NHST will beget a corresponding practice with confidence intervals. However, just because a confidence interval can be interpreted in some ways like a statistical test does not mean that it must be.

Confidence intervals are not a magical alternative to NHST. However, interval estimation in individual studies and replication together offer a much more scientifically credible way to address sampling error than the use of statistical tests in individual studies. Consider again the data in Table 3.2 for six hypothetical replications. A 95% confidence interval about the observed mean contrast is reported for each replication. Each interval estimates sampling error, but itself is also subject to sampling error. The range of overlap among the six confidence intervals is 2.06 to 5.03. This information is more useful than knowing that a nil hypothesis was rejected in 3/6 studies. F. Schmidt (1996) and others have noted that even if our initial expectations regarding parameters are very wrong, we will eventually discover our error by plotting the related confidence intervals across studies.

2. *Misinterpretations of NHST are not the fault of the method.* Defenders of NHST generally acknowledge widespread misinterpretations. They also note that such misunderstandings are the responsibility of those who use it (Krantz, 1999). Critics may counter that any method with so much apparent potential to be misconstrued by so many intelligent and highly educated users must ultimately assume some of the blame.
3. *More careful use of technical terms may avoid unwarranted connotations.* An area of suggested reform concerns the language used to report the results of statistical tests (e.g., D. Robinson & Levin, 1997). For example, some have suggested that the term *significant* should always be qualified by the word *statistically*—which may prompt readers to distinguish between statistical significance and substantive significance (B. Thompson, 1996)—and that exact p values should be reported instead of just whether they are less than or greater than α, such as:

$$t\,(20) = 2.40,\ p = .026$$

instead of

$$t\,(20) = 2.40,\ p < .05$$

The latter recommendation has some problems, however. The possibility that p values are incorrect in many behavioral studies was mentioned earlier, so their reporting to three- or four-decimal accuracy may give a false impression. In large samples, p values are often very low, such as .000012, and reporting such small probabilities may actually encourage misinterpretation. It must also be said that these kinds of suggestions have been made many times over the past 50 years with little apparent impact. Critics would probably feel little conviction that any of the modifications just described would ameliorate the limitations of NHST for most applications in the social sciences. For them, the following expression may be pertinent: You can put candles in a cow pie, but that does not make it a birthday cake.

4. *Some research questions require a dichotomous answer.* The final outcome of NHST is the decision to reject or fail to reject H_0. There are times when the question that motivates the research is also dichotomous, including, for instance, should this intervention program be implemented? Is this drug more effective than placebo? The method of NHST addresses whether observed effects or relations stand out above sampling error, but it is not as useful for estimating the magnitudes of these effects (Chow, 1996). There are also times when theories predict directions of effects but not their specific magnitudes. One reading instruction method may be believed to be more effective than another by some unknown amount, for example. The testing of theories that predict directions but not amounts is also probably more typical in the social sciences than in the natural sciences. However, it is always useful to measure the magnitude of an effect. Indeed, if we cannot think about magnitudes, then we may never get to theories that predict magnitudes instead of just directions. Estimating the average size of an effect with meta-analysis instead of counting the numbers of H_0 rejections is also a better way to synthesize results across a set of studies (chap. 8, this volume).
5. *Nil hypotheses are sometimes appropriate.* The criticism that nil hypotheses are typically false was discussed earlier. As noted by Frick (1995), D. Robinson and Wainer (2002), and others, there are cases when the assumption of a zero effect is justified. For example, there may be no reason in a complex study to predict an effect when just one independent variable is manipulated.

6. *The method of NHST is a gateway to statistical decision (utility) theory*. In this approach—well known in fields such as engineering and environmental studies—probabilities of Type I and Type II errors are weighted by estimated costs of each kind of mistake. The net anticipated gains and losses are then evaluated to make rational decisions about alternative actions in the face of uncertainty. In contrast to NHST, the probability of a Type I error is not arbitrarily set to either .05 or .01 in statistical decision theory. The latter method may be able to detect long-term negative consequences of an intervention even while statistical tests are unable to reject the nil hypothesis of no short-term effect (Johnson, 1999). Statistical decision theory is a very powerful method if it is possible to estimate the costs of different decisions in dollars, life expectancy, or some other quantitative, objective metric. This is not usually possible in behavioral research.

VARIATIONS ON NHST

This section identifies some specialized methods that are modifications of the basic NHST model. These methods may avoid some of the problems of traditional statistical tests and can be very useful in the right situation. It is possible to give only brief descriptions, but interested readers can look to the works cited next for more information.

Range Null Hypotheses and Good-Enough Belts

As mentioned, any point null hypothesis is probably false. Serlin (1993) described the specification of H_0 as a range hypothesis that indicates the values of the population parameter considered equivalent and uninteresting. The alternative hypothesis is still a range hypothesis, but it specifies a minimum result based on substantive considerations that is necessary for additional analysis. These ranges for H_0 and H_1 are called *good-enough belts*, which implies that one hypothesis or the other is considered supported within predefined margins. The specification of range null hypotheses in the social sciences is relatively rare—a notable exception is the evaluation of model fit in structural equation modeling (e.g., Kaplan, 2000, chap. 6)—and there is some question whether this approach would make any practical difference (Cortina & Dunlap, 1997).

Equivalence Testing

Equivalence testing is better known in pharmacological research and the environmental and biological sciences. It deals with the problem of establishing equivalence between two groups or conditions. For example, a researcher may wish to determine whether a generic drug can be substituted for a more expensive drug. In traditional NHST, the failure to reject H_0: $\mu_1 = \mu_2$ is not evidence that the drugs are equivalent. In one form of equivalence testing, a single point null hypothesis is replaced by two range subhypotheses. Each subhypothesis expresses a range of $\mu_1 - \mu_2$ values that corresponds to substantive mean differences. For example, the pair of subhypotheses

$$H_0: \begin{cases} H_{0_1}: (\mu_1 - \mu_2) < -10.00 \\ H_{0_2}: (\mu_1 - \mu_2) > 10.00 \end{cases}$$

says that the population means cannot be considered equivalent if the absolute value of their difference is greater than 10.00. The complementary interval for this example is the equivalence hypothesis

$$-10.00 \leq (\mu_1 - \mu_2) \leq 10.00$$

which is a good-enough belt for the equivalence hypothesis. Standard statistical tests are used to contrast the observed mean difference against each of these one-sided null hypotheses for a directional alternative hypothesis. Only if *both* range subhypotheses are rejected at the same level of α can the compound null hypothesis of nonequivalence be rejected. The same decision can also be reached on the basis of a confidence interval around the observed mean difference. In the approach just outlined, Type I error is the probability of declaring two populations or conditions to be equivalent when in truth they are not. In a drug study, this risk is the patient's (consumer's) risk. McBride (1999) showed that if Type I error risk is to be the producer's instead of the patient's, the null hypothesis appropriate for this example would be

$$H_0: -10.00 \leq (\mu_1 - \mu_2) \leq 10.00$$

and it would be rejected if either the lower end of a one-sided confidence interval about the observed mean difference is greater than 10.00 or the upper end of a one-sided confidence interval is less than −10.00. Rogers, K. Howard, and Vessey (1993) introduced equivalence testing to social

scientists, and P. M. Dixon (1998) described its application in risk assessment.

Inferential Confidence Intervals

Tryon (2001) proposed an integrated approach to testing means for statistical difference, equivalence, or indeterminancy (neither statistically different or equivalent). It is based on *inferential confidence intervals*, which are modified confidence intervals constructed around individual means. The width of an inferential confidence interval is the product of the standard error of the mean (Equation 2.4) and a two-tailed critical t value reduced by a correction factor that equals the ratio of the standard error of the mean difference (Equation 2.8) over the sum of the individual standard errors. Because values of this correction factor range from about .70 to 1.00, widths of inferential confidence intervals are generally narrower than those of standard confidence intervals about the same means.

A statistical difference between two means occurs in this approach when their inferential confidence intervals do not overlap. The probability associated with this statistical difference is the same as that from the standard t test for a nil hypothesis and a nondirectional alternative hypothesis. In other words, this method does not lead to a different conclusion than standard NHST, at least in difference testing. Statistical equivalence is concluded when the *maximum probable difference* between two means is less than an amount considered inconsequential as per an equivalence hypothesis. The maximum probable difference is the difference between the highest upper bound and the lowest lower bound of two inferential confidence intervals. For example, if the 10.00 to 14.00 and 12.00 to 18.00 are the inferential confidence intervals based on two different means, the maximum probable difference is $18.00 - 10.00 = 8.00$. If this difference lies within the range set by the equivalence hypothesis, statistical equivalence is inferred. A contrast neither statistically different or equivalent is indeterminant, and it is not evidence for or against any hypothesis. Tryon claimed that this method is less susceptible to misinterpretation because (a) the null hypothesis is implicit instead of explicit, (b) the model covers tests for both differences and equivalence, and (c) the availability of a third outcome—statistical indeterminancy—may reduce the interpretation of marginally nonsignificant differences as "trends." It remains to be seen whether this approach will have a positive impact.

Three-Valued Logic

Kaiser (1960) may have been one of the first social science authors to suggest substituting *three-valued logic* for the standard two-valued (dichotomous) logic of NHST. Briefly, three-valued logic allows split-tailed alternative hypotheses that permit statistically significant evidence against a substantive hypothesis if the direction of the observed effect is not as predicted. This is basically a simultaneous test of two directional alternative hypotheses, one for and the other against the research hypothesis. The third kind of test is for a standard nondirectional alternative hypothesis. Harris (1997b) provides a very clear, contemporary description of three-valued logic, but notes that it has not been used much.

WHAT DO WE DO NOW? BUILDING A BETTER FUTURE

After considering criticisms of statistical tests, we can choose one of the following courses of action:

1. Do nothing; that is, continue using statistical tests just as we have for the past 50 years.
2. Stop using statistical tests entirely. Stop teaching them in university courses. Effectively ban their use by refusing to publish studies in which they are used. Although this option sounds hypothetical or even radical, some highly respected researchers have called for such a ban (e.g., Hunter, 1997; F. Schmidt, 1996).
3. Chart a course between the two extremes listed, one that calls for varying degrees of use of statistical tests—from none to somewhat more pivotal, depending on the research context, but with strict requirements for their use.

The first option is not acceptable because there are negative implications for the advancement of behavioral research that rule out doing nothing. A ban on statistical tests in psychology journals does not seem imminent in the short term, but the fact that some journals require the reporting of effect sizes is an effective ban on the use of statistical tests by themselves (see chap. 1, this volume). The first two options are thus excluded.

Outlined next are recommendations based on the third option. They are intended as a constructive framework for change. It is assumed that reasonable people could disagree with some of the specifics put forward. Indeed, a lack of consensus has characterized the whole debate about NHST, so no single set of recommendations will satisfy everyone. Even if the reader does not endorse all the points outlined, he or she may at least learn new ways of looking at the controversy about statistical tests or, even better, data, which is the ultimate goal of this book.

The main theme of the recommendations can be summarized like this: The method of NHST may have helped us in psychology and related

behavioral sciences through a difficult adolescence during which we struggled to differentiate ourselves from the humanities while at the same time we strived to become more like our primary role model, the natural sciences. However, just as few adults wear the same style of clothes, listen to the same types of music, or have the same values they did as teenagers, behavioral research needs to leave its adolescence behind and grow into new ways of doing things. Arrested development is the only alternative.

Recommendations

Specific suggestions are listed and then discussed afterward:

1. Only in very exploratory research where it is unknown whether effects exist may a primary role for NHST be appropriate.
2. If statistical tests are used, (a) information about power must be reported, and (b) the null hypothesis must be plausible.
3. In any kind of behavioral research, it is not acceptable anymore to describe results solely in terms of NHST outcomes.
4. Drop the word "significant" from our data analysis vocabulary. Use it only in its everyday sense to describe something actually noteworthy or important.
5. It is the researcher's responsibility to report and interpret, whenever possible, effect size estimates and confidence intervals for primary results. This does *not* mean to report effect sizes only for H_0 rejections.
6. It is also the researcher's responsibility to demonstrate the substantive (theoretical, clinical, or practical) significance of the results. Statistical tests are inadequate for this purpose.
7. Replication is the best way to deal with sampling error.
8. Education in statistical methods needs to be reformed, too. The role of NHST should be greatly deemphasized so that more time can be spent showing students how to determine whether a result has substantive significance and how to replicate it.
9. Researchers need more help from their statistical software to compute effect sizes and confidence intervals.

A Primary Role for NHST May Be Suitable Only in Very Exploratory Research

The ability of NHST to address the dichotomous question of whether relations are greater than expected levels of sampling error may be useful in some new research areas. Considering the many limitations of NHST discussed, the period of this usefulness should be brief. Given evidence that

an effect exists, the next steps should involve estimation of its magnitude and evaluation of its substantive significance, both of which are beyond the range of what NHST can tell us. More advanced study of the effect may require model-fitting techniques, such as structural equation modeling (e.g., Kline, 1998), hierarchical linear modeling (e.g., Raudenbush & Bryk, 2002), or latent class analysis (e.g., Hagenaars & McCutcheon, 2002), among other techniques that test models instead of just individual hypotheses. It should be the hallmark of a maturing research area that NHST is not the primary inference tool.

Report Power for Any Use of Statistical Tests, and Test Only Plausible Null Hypotheses

The level of power reported should be a priori power, not observed power (see chap. 2, this volume; also Wilkinson & TFSI, 1999). It is especially important to report power if most of the results are negative—that is, there were few H_0 rejections. This is because readers of an empirical study should be able to tell whether the power of the study is so low (e.g., < .50) that negative results are expected. Knowing that H_0 was rejected is useful only if that hypothesis is plausible. Also, one minus power is the probability of a Type II error, which can only occur if H_0 is not rejected when there is a real effect. We probably see so few examples of reporting power in the research literature when the results are mainly negative because of bias toward only publishing studies with H_0 rejections. In a less biased literature, however, information about power would be more relevant. Low p values that exaggerate the relative infrequency of the results are expected under implausible null hypotheses. If it is feasible to test only a nil hypotheses but a nil hypothesis is implausible, interpretation of the results of statistical tests should be accordingly modified.

It Is Not Acceptable to Describe Results Only on the Basis of NHST Outcomes

All of the shortcomings of NHST considered provide the rationale for this recommendation. For journal editors and reviewers, NHST outcomes should also not be the primary consideration for deciding whether to accept or reject submissions for publication.

Stop Using the Word "Significant" in Connection With NHST

In hindsight, the choice of the word "significant" to describe the event $p < \alpha$ was very poor. Although statisticians understand that significant in NHST does *not* imply a large or important effect, the use of this word may foster false beliefs among nonstatisticians. Accordingly, we in the behavioral sciences should "give" this word back to the general public and use it only

as does everyone else—to denote importance, meaningfulness, or substantialness. Use of just the word "statistical" when H_0 is rejected should suffice. For instance, rejection of the hypothesis H_0: $\rho = 0$ could be described as evidence for a statistical association or correlation between the variables, and rejection of the hypothesis H_0: $\mu_1 = \mu_2$ could be described as evidence for a statistical mean difference (Tryon, 2001). Calling an effect statistical implies that it was observed, but not also noteworthy. Of course, statistical effects may also be meaningful effects, but this is a not a question for NHST. The simple phrasing just suggested also seems preferable to the expression "statistically reliable" to describe H_0 rejections. This is because one connotation of reliable is repeatable, but p values say nothing directly about the chances of replication. At the very least, if the word significant is used in an oral or written interpretation of the results, it should *always* be preceded by the qualifier "statistical" (B. Thompson, 1996).

Whenever Possible, Researchers Should Be Obliged to Report and Interpret Effect Sizes and Confidence Intervals

That increasing numbers of journals require effect size estimates supports this recommendation (see chap. 1, this volume). Reporting confidence intervals for effect size estimates is even better: Not only does the width of the confidence interval directly indicate the amount of sampling error associated with an observed effect size, it also estimates a range of effect sizes in the population that may have given rise to the observed result. However, it is recognized that it is not always possible to compute effect sizes in certain kinds of complex designs or construct confidence intervals based on some types of statistics. However, effect size can be estimated and confidence intervals can be reported in most behavioral studies.

Researchers Should Also Be Obliged to Demonstrate the Substantive Significance of Their Results

Null hypothesis rejections do not imply substantive significance. Thus, researchers need other frames of reference to explain to their audiences why the results are interesting or important. A quick example illustrates this idea. In a hypothetical study titled "Smiling and Touching Behaviors of Adolescents in Fast Food Restaurants," effects were statistically significant, but might not be deemed substantively important to many of us, especially if we are not adolescents, or do not frequent fast food restaurants. It is not easy to demonstrate substantive significance, and certainly much more difficult than using p values as the coin of the social scientific realm. Estimation of effect size gives a starting point for determining substantive significance; so does consulting meta-analytic works in the area (if they exist). It is even better for researchers to use their domain knowledge to inform the

use of the methods just mentioned (Kirk, 1996). These ideas are elaborated in the next chapter.

Replication Is the Best Way to Deal With Sampling Error

The rationale for this recommendation is also obvious. It would make a very strong statement if journals or granting agencies required replication. This would increase the demands on the researcher and result in fewer published studies. The quality of what would be published might improve, however. A requirement for replication would also filter out some of the fad social science research topics that bloom for a short time but then quickly disappear. Such a requirement could be relaxed for original results with the potential for a large impact in their field, but the need to replicate studies with unexpected or surprising results is even greater (D. Robinson & Levin, 1997).

Education in Statistical Methods Should Be Much Less NHST-Centric and More Concerned With Replication and Determining Substantive Significance

The method of NHST is often presented as the pinnacle in many introductory statistics courses. This situation is reinforced by the virtually monolithic, NHST-based presentation in many undergraduate-level statistics textbooks. Graduate courses often do little more than inform students about additional NHST methods and strategies for their use (Aiken et al., 1990). The situation is little better in undergraduate psychology programs, which emphasize traditional approaches to analysis (i.e., statistical tests) and have not generally kept pace with changes in the field (Frederich, Buday, & Kerr, 2000). It is also true that many statistics textbooks still do not emphasize methods beyond traditional statistical tests, such as effect size (e.g., R. Capraro & M. Capraro, 2002).

Some topics already taught in introductory courses should be given more prominence. Many effect size indexes are nothing more than correlations, proportions of standard deviations, or percentages of scores that fall at certain points. These are all basic kinds of statistics covered in many introductory courses. However, their potential application outside classical descriptive or inferential statistics is often unexplained. For example, students usually learn about the t test for comparing independent means. These same students often do not know about the point-biserial correlation, r_{pb}. In a two-sample design, r_{pb} is the correlation between a dichotomous independent variable (group membership) and a quantitative dependent variable. It is easily derived from the t test and is just a special case of the Pearson correlation r.

Less emphasis on NHST may also encourage students to choose simpler methods of data analysis (e.g., Wilkinson et al., 1999). Doing so may help

them appreciate that NHST is *not* necessary to detect meaningful or noteworthy effects, which should be obvious to visual inspection of relatively simple kinds of statistics or graphical displays (J. Cohen, 1994). The description of results at a level closer to the data may also help students develop better communication skills. This is important for students who later take up careers where they must explain the implications of their results to policy makers (McCartney & R. Rosenthal, 2000).

We also need better integration between courses in research methods and statistics. In many undergraduate programs, these subjects are taught in separate courses, and there may be little connection between the two. The consequence is that students learn about data analysis methods without getting a good sense of their potential applications. This may be an apt time to rethink the partition of the teaching of research skills into separate statistics and methods courses.

Statistical Software Should Be Better at Computing Effect Sizes and Confidence Intervals

Most general statistical software programs are still very NHST-centric. That more of them now optionally print at least some kinds of effect size indexes is encouraging. Considering these discussions, however, perhaps results of statistical tests should be the optional output. Literally dozens of effect size indexes are available (e.g., Kirk, 1996), and at least some of the more widely used indexes should be available in computer program output for every analytical choice. It should also be the case that for a given analytical choice, several different effect sizes are options. Many contemporary general statistical programs also optionally print confidence intervals for population means or regression coefficients, but they should also give confidence intervals for population effect sizes. It was mentioned that special computational methods required for exact confidence intervals for effect sizes have only recently become more widely available, but one hopes that these algorithms will soon be incorporated into general statistical packages.

CONCLUSION

Statistical tests have been like a collective Rorschach inkblot test for the social sciences: What we see in them has had more to do with wish fulfillment than what is really there. This collective magical thinking has impeded the development of psychology (and related areas) as a cumulative science. There is also a mismatch between the characteristics of many behavioral studies and what is required for results of statistical tests to be accurate. That is, if we routinely specified plausible null hypotheses, studied random samples, checked distributional assumptions, estimated power, and understood the correct meaning of p values, there would be no problem with statistical tests as our primary inference tool. *None* of these conditions are generally true in the behavioral sciences. I offered several suggestions in this chapter, all of which involve a much smaller role—including none whatsoever—for traditional statistical tests. Some of these suggestions include the computation of effect sizes and confidence intervals for all effects of interest, not just ones in which a null hypothesis is rejected, and evaluation of the substantive significance of results, not just their statistical significance. Replication is the most important reform of all.

RECOMMENDED READINGS

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

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

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: Erlbaum.