Psychological Methods
1996, Vol. 1, No. 4, 379-390

1082-989X/96/$3.00

The Appropriate Use of Null Hypothesis Testing

Robert W. Frick
State University of New York at Stony Brook

The many criticisms of null hypothesis testing suggest when it is not useful and what is should not be used for. This article explores when and why its use is appropriate. Null hypothesis testing is insufficient when size of effect is important, but it is ideal for testing ordinal claims relating the order of conditions, which are common in psychology. Null hypothesis testing also is insufficient for determining beliefs, but it is ideal for demonstrating sufficient evidential strength to support an ordinal claim, with sufficient evidence being 1 criterion for a finding entering the corpus of legitimate findings in psychology. The line between sufficient and insufficient evidence is currently set at $p < .05$; there is little reason for allowing experimenters to select their own value of alpha. Thus null hypothesis testing is an optimal method for demonstrating sufficient evidence for an ordinal claim.

Null hypothesis testing is a statistical procedure used by most experimenters. Roughly speaking, an experimenter constructs a null hypothesis, such as there is no difference between conditions or no association between two variables, then calculates a value of p, which is conventionally defined as the probability of achieving the observed outcome or larger, given the null hypothesis. When p is less than some criterion, which is almost always .05, the experimenter "rejects" the null hypothesis and concludes that (a) one condition is better than another, (b) there is an association between two variables, or (c) some particular pattern exists in the data (Johnstone, 1987; Kaiser, 1960).

Null hypothesis testing has received severe criticism; among many it is taken as obvious that it should be abandoned (Cohen, 1994). For example, Oakes (1986, p. vii) wrote, "Many researchers retain an infatuation with significance tests despite the formidable arguments that have been presented against them. In Chapters 1–3 I marshal these arguments . . . in an attempt to kill the beast—but I suspect the headless corpse will continue to flail through journal pages for years to come." Or, "There is a long and honorable tradition of blistering attacks on the role of significance testing in the behavioral sciences, a tradition reminiscent of knights in shining armor bravely marching off, one by one, to slay a rather large and stubborn dragon. . . . Given the cogency, vehemence and repetition of such attacks, it is surprising to see that the dragon will not stay dead" (Harris, 1991, p. 375).

Despite these attacks, null hypothesis testing still dominates the social sciences (Loftus & Masson, 1994). Its continued use is typically attributed to experimenters' ignorance, misunderstanding, laziness, or adherence to tradition (Falk & Greenbaum, 1995; Johnstone, 1988; Nunnally, 1960; Oakes, 1986; Weitzman, 1984). However, as an anonymous reviewer put it, "A way of thinking that has survived decades of ferocious attacks is likely to have some value."

This article explains the value of null hypothesis testing. The attacks on null hypothesis testing point out its limitations but do not rule out its appropriate use. This article also builds on previous attempts to defend null hypothesis testing (e.g., Chow, 1988, 1991; Cox, 1977; Giere, 1972; Greenwald, Gonzalez, Harris, & Guthrie, 1996; Kalbfleisch & Sprott, 1976).

First, this article will consider the issue of the type of claim experimenters should be making, comparing what I will call *quantitative* to *ordinal* claims. For quantitative claims, null hypothesis testing is not sufficient and perhaps not the statistic of choice, but for ordinal claims it is ideal. Next, this article considers the role of null hypothesis testing in establishing these

I thank Art Aron, Darla Broberg, Dave Cross, and anonymous reviewers for commenting on draft versions of this article and the members of edstat-l for useful discussions.

Correspondence concerning this article should be addressed to Robert W. Frick, Department of Psychology, State University of New York, Stony Brook, New York 11790-2500. Electronic mail may be sent via Internet to rfrick@psych1.psy.sunysb.edu.

ordinal claims. Null hypothesis testing is not sufficient for establishing beliefs or estimating the probability that these ordinal claims are correct. Instead, it establishes that sufficient evidence has been presented to support a claim, with *sufficient* defined as $p < .05$. This is one criterion for a finding entering the corpus of psychology. Finally, reasons are presented for why experimenters are not allowed to choose their own value of alpha.

This article will be concerned with the use of null hypothesis testing when an effect is found, which is to say, when statistical significance is achieved. It is well agreed that null hypothesis testing by itself does not provide sufficient evidence for accepting the null hypothesis. (The issue of accepting the null hypothesis is addressed in Frick, 1995.)

When there is an effect, but the experiment does not have sufficient power to detect that effect using null hypothesis testing, the outcome is unfortunate. However, the problem is insufficient power, not the use of null hypothesis testing, and the only solution is to increase power (e.g., test more subjects or perform a meta-analysis across experiments).

Supporting Ordinal Claims

One common criticism of null hypothesis testing begins with the assumption that experimenters should be interested in the size of an effect. If a drug impairs IQ 6 points, it is natural and appropriate to think of 6 as being a measure of "effect size" (Cohen, 1988, p. 10; Hunter & Schmidt, 1990, p. 233; Richardson, 1996), though there are a variety of other measures of effect size (cf. the American Psychological Association's *Publication Manual*, 1994, p. 18; Richardson, 1996). I call all claims about size of effect quantitative claims.

The assertions that conventional statistical testing actually slows or impedes science (Cohen, 1994; Loftus, 1994; Oakes, 1986) probably are based on the belief that estimating effect size is the ultimate goal of science. Cohen (1988) wrote, "A moment's thought suggests that it [effect size] is, after all, what science is all about. For sure, it's not about significance testing" (p. 532). If the goal of an experiment was to make a claim concerning size of effect, one could simply report the observed size of effect. If it was desirable to express the precision of this estimate, a confidence interval around this observed size of effect could also be reported. Null hypothesis testing would not be needed, and in fact would draw attention away from the size of the effect.

Instead of claiming the drug increases IQ 6 points, one could claim merely that the drug improves IQ. For convenience in exposition, I will refer to this type of claim as ordinal. An *ordinal claim* can be defined as one that does not specify the size of effect; alternatively, it could be defined as a claim that specifies only the order of conditions, the order of effects, or the direction of a correlation. The statistical operations used to justify these claims usually assume more than an ordinal scale, so an ordinal claim is ordinal only in the sense that the final claim is about order. In a factorial design, the ordinal claim of an interaction would be, for example, that an effect is larger in one situation than in another, without specifying the size of the difference between the two situations. In a correlational study, for example, obtaining a correlation between smoking and lung cancer, the value of r is a measure of the size of effect and the ordinal claim would be that smoking has a positive correlation with lung cancer.

Null hypothesis testing is used to provide support for ordinal claims, because establishing a pattern of order requires ruling out equivalence. The problem is that an observed effect in the data could have been caused by chance fluctuations, not by some "real" effect. Null hypothesis testing addresses whether or not there is sufficient evidence to support the existence of an effect. For example, to claim that a drug improved IQ, the experimenter would need statistical significance for the null hypothesis that the drug had no effect on IQ; to claim a correlation between smoking and lung cancer, the experimenter would need statistical significance for the null hypothesis that $r = 0$. This explains the common use of 0 as a null hypothesis: The value of 0 neatly divides the space of possible effect sizes into the three relevant categories of no effect (the drug does not influence IQ), effect in one direction (the drug increases IQ), and effect in the other direction (the drug impairs IQ; Cox, 1977).

Knowing the exact size of effect would imply the direction of effect. However, no experiment ever establishes the exact size of an effect; all the experiment can do is establish the approximate size of effect, and knowing the approximate size of effect does not necessarily establish the direction of effect. Therefore, the two are different goals. For example, consider the confidence interval for a drug's effect on IQ. An experiment producing a confidence interval of (−3, 7) establishes the size of effect just as well as an experiment producing a confidence interval of (10, 20), but only the latter establishes the ordinal claim that the drug improves IQ.

A critical issue for the value of null hypothesis testing, hence, is whether the goal of experimenting always is determining the size of effect. I will consider three different statistical niches, formed by three different uses for the finding of an experiment: (a) testing a model making quantitative predictions, (b) supporting or disconfirming a law or theory, and (c) directly applying the finding to a practical situation. The argument will be that for one of these niches null hypothesis testing is inappropriate, for one it is insufficient but potentially useful, and for one it is ideal.

Models Making Quantitative Predictions

It is useful to think of theories as being statements about the workings of an underlying reality. As such, theories do not make quantitative predictions about the values that will be observed in the world. For example, Newton's theory includes statements about the force of gravity and the conservation of energy and momentum. Newton's theory, by itself, makes no predictions about the path of Mars (or any other object). However, with estimates of the mass of the sun and Mars, and the current position and velocity of Mars (with respect to the sun), Newton's theory can be used to model the path of Mars around the sun. Thus, models can be built with the purpose of making quantitative predictions. Models can also be based on regression estimates of potentially relevant variables.

For the purpose of testing the model, the issue might seem to be the accuracy of its predictions. However, as is well-accepted, quantitative predictions are never perfectly accurate. One problem is that the input variables are rarely perfectly accurate. A second problem is that a model usually ignores factors that might be having small effects. Third, when a model is based on regression estimates of potentially relevant variables, it might assume linear relationships, which is rarely correct.

For example, the model of the path of Mars presented here will necessarily be wrong, independent of any problems in Newton's theory. First, the estimates concerning the Sun and Mars will not be perfectly accurate. Second, this model ignores the influence of other heavenly bodies. The fact that the model is not perfect is hence uninteresting and does not disconfirm Newton's theory.

Null hypothesis testing could be used to compare the model's prediction to the observed value. However, because the fallibility of the model is known in advance, a statistically significant discrepancy from the model's prediction is not informative. This test will always be statistically significant, given enough power (Grant, 1962). Conversely, a lack of statistical significance does not mean the model is correct, it just means that not enough observations were made.

Thus, null hypothesis testing is not useful for testing a model (Berkson, 1938; Grant, 1962). Similarly, in comparing two models, it would be inappropriate to conclude that one model is correct just because another has been shown to be wrong by statistical testing.

Testing Laws and Theories

A second statistical niche is formed by experiments supporting or disconfirming a law or theory. To start, consider laws, which can be defined as claims relating the changes in two (or more) variables. A law is supported by demonstrating that it holds in a particular experiment.

An examination of any textbook will reveal that the laws in psychology are usually stated in an ordinal form, not quantitatively. For example, one law in psychology is that frustration increases the tendency to aggression. To test this law experimentally, the question for statistics would be whether frustration increased the amount of observed aggression, with the size of increase being irrelevant. Null hypothesis testing is ideal for supporting these ordinal laws. The size of the effect is irrelevant for supporting the law, and scientists seem to be interested in laws whether the effect size is large or small.

The size of effect is not completely irrelevant. First, the size of effect might signal the importance of the manipulated variable in the process being studied. Second, the robustness of the effect is important for purposes of further experimentation, with larger effect sizes being easier to replicate. However, for these two purposes, the exact size of effect is not important. For example, a classic experiment by Sperling (1960) compared whole report to partial report for brief visual presentations. Whole report capacity was about 4.5 and partial report suggested a capacity of about 9. It is important that the difference between these two conditions was 4.5 rather than .5. However, a difference of 3.5 or 13.5 would have had essentially the same meaning.

The point of disagreement with regard to null hypothesis testing is not whether the current laws in psychology are ordinal; the disagreement is whether

experiments should be trying to support ordinal laws. With everything else being equal, a quantitative law contains more information than an ordinal law and hence would be preferred.

The problem for quantitative laws, however, is generality. It is easy to form a quantitative law on the basis of an experimental finding—one simply reports the size of the effect. However, a law holding for just one very narrow situation is not very useful and cannot be usefully added to the collection of laws comprising the science of psychology. Therefore, for a law to be of any value, it must generalize across different situations.

As examples will suggest, there are substantial and perhaps insurmountable obstacles to forming quantitative laws with generality. Consider the law that frustration increases the tendency to aggression. The exact size of the effect would depend on a very large number of different factors. First, there are a variety of different ways of measuring aggression. Second, the amount of increase in aggression depends on the size of the manipulation of frustration. These are a variety of ways of manipulating frustration, and the effectiveness of each will vary. The ordinal law would be expected to generalize to new methods of producing frustration and new methods of measuring tendency to aggression; it is difficult to imagine how any quantitative expression of this law would generalize to new manipulations of frustration or new measures of aggression.

Third, it would be unlikely that any scale of aggression would be an interval scale, which would be needed to make quantitative predictions with any economy. For example, if a given amount of frustration is going to increase aggression by 2, it must increase the level of aggression from 1 to 3, from 2 to 4, and from 3 to 5, which implies an interval scale. With just an ordinal scale, there is no simple relationship between increase in frustration and increase in aggression—instead, the relationship between aggression and frustration would have to be plotted for each level of frustration.

Fourth, the ordinal law that frustration increases aggression presumably applies to a wide variety of people and situations. However, a law claiming a specific size of increase in aggression will apply only to the sample and situation used in the experiment supporting that law. For example, the effect might be larger in less emotionally mature people, or it might depend on intelligence, age, or amount of schooling. Sears (1986) noted that the size of effect for social phenomena in college students does not generalize well to noncollege students, but the ordinal pattern does.

Thus, examples suggest several reasons why psychology does not contain quantitative laws with generality. One exception is in psychophysics. Psychophysicists have interval scales for their manipulations (e.g., brightness), they use a common scale for measurement (e.g., probability of detection), and they try to control for the effect of all extraneous variables (such as amount of adaption to light). As a result, they can and do attempt to construct quantitative conclusions, with Weber's Law standing as a good example. Nonetheless, examination of a sensation and perception textbook (Levine & Shefner, 1991) reveals a presentation of more ordinal than quantitative claims, suggesting a small range for quantitative laws.

The story is the same for theories. Theories can be roughly defined as one or more statements (a) about underlying and often unobservable constructs that (b) together make predictions. Currently, most theories in psychology yield predictions about ordinal patterns, not size of effect. For example, the existence of iconic memory predicts that, for brief visual presentations, partial report can be superior to full report. Thus, testing theories requires testing ordinal predictions, so null hypothesis testing is ideal for this niche.

Physics has had great success with theories that make quantitative predictions, and many people hope that the theories in psychology will eventually make quantitative predictions. However, again there seems to be possibly insurmountable problems in constructing theories that by themselves make quantitative predictions. Consider the prediction of better partial report of a brief visual presentation. The existence of iconic memory makes no prediction about the size of this effect, because the size of effect depends on many factors completely extraneous to the theory. First, the size of increase depends on details concerning the stimulus array. Second, it depends on the rate of decay from iconic memory versus the time at which subjects process the cue and focus on the relevant part of the display. For example, Sperling used tones to cue subjects as to which row would be reported; subject's performance was influenced by how perceivable and discriminable the tones were.

Therefore, prior to performing the experiment, it would have been impossible to make a quantitative prediction. After the experiment was performed, a quantitative prediction would be possible for the particular situation tested in the experiment. However,

this would not be an example of the theory making a quantitative prediction.

Sometimes a theory will predict that one effect is larger than another. This is a prediction about effect size, but it is still an ordinal prediction: Is there a difference in effect sizes, and if so, which is larger? Similarly, with regard to laws, one can ask if an effect in one situation is the same, larger, or smaller in a new situation. Conventional statistical testing can address the comparison of sizes of effect in two ways. If the exact size of the original effect was already known, it could be used as the null hypothesis. If the size of the first effect is also being measured, an F test can be performed for the presence of an interaction, comparing the effect size in one condition to the effect size in another condition.

Thus, the current status of psychology is that theories and laws are tested by ordinal patterns. Null hypothesis testing is then used to support these laws and theories. A goal of many is that the laws and theories in psychology be quantitative. Physics obviously has had great success with quantitative laws, so it is natural to hope that psychology could have the same success. However, the goal of quantitative laws is old, and it has not proved very successful. Some people are still working toward this goal, and perhaps they will eventually succeed. However, it is at least possible that the long-standing goal to found psychology as a quantitative science might be impossible and that psychology perhaps should take pride in its successes as an ordinal science. In any case, current experimenters have to deal with the current condition, which is laws and theories making ordinal predictions.

Practical Applications

The third statistical niche is formed by experiments intending to use the results for immediate practical application. For example, an experiment might find that Treatment A is more effective than Treatment B, then use this finding to support the use of Treatment A.

If everything else is constant, then this ordinal claim is enough to decide that Treatment A should be used over Treatment B. However, it is rare for everything else to be held constant. For example, the treatments might differ in their cost or side effects. Whenever a decision must be made balancing costs and benefits, it is important to consider the size of the costs and benefits. Therefore, the size of effect is always important in the practical application experiment.

Because size of effect is important, exclusive reliance on null hypothesis testing is inappropriate. Instead, it is important to distinguish statistical significance from "clinical significance" or "practical significance" (Cohen, 1965; Edwards, 1950; Grant, 1962; Tyler, 1931)—just because a result is statistically significant does not mean that it is clinically significant.

However, although null hypothesis testing is insufficient, that does not mean that null hypothesis testing cannot play a role. Some experimenters apparently use null hypothesis testing in the following manner. First, a difference is shown to be statistically significant. Then the obtained effect size is used as an estimate of the actual effect size. Finally, either the experimenter or consumer judges whether or not the obtained effect size is clinically significant. Another possibility is to first determine the minimal size of a clinically significant effect. This minimally relevant size can be used as a null hypothesis. Rejecting this null hypothesis would suggest that the true difference is clinically significant (Fowler, 1985).

Confidence Intervals

Two different functions of confidence intervals should be distinguished. First, confidence intervals can be used to indicate the precision of the estimate of effect size. For example, the 95% confidence interval of 6 ± 2 shows more precision than the 95% confidence interval of 6 ± 5. This function of the confidence interval is appropriate whenever size of effect is the issue. Technically, whether or not 0 is part of the confidence interval is not an issue.

Second, a 95% confidence interval can function to indicate which values could not be rejected by a two-tailed test with alpha at .05. In this function, the confidence interval could replace the report of null hypothesis for just one value, instead communicating the outcome of the tests of all values as null hypotheses (e.g., Cohen, 1994). This function does not avoid the logic of null hypothesis testing, however—Cohen (1994) was being illogical when he criticized the logic of null hypothesis testing and then advocated using the confidence interval because it reported the results of all statistical tests.

Usually there is one particular value of interest as a null hypothesis, because rejecting this value establishes an ordinal pattern the experimenter wishes to claim. The value of p for this null hypothesis contains important information, as is discussed later in this article, so it should be reported (Greenwald et al.,

1996). It would be inappropriate to replace this report of p with a confidence interval (Spjotvoll, 1977). Therefore, replacing the report of the test of one null hypothesis with a confidence interval is inappropriate when null hypothesis testing is being used to support an ordinal claim.

Evaluation

As critics have long noted, not all experimenters can mindlessly perform null hypothesis testing, achieve statistical significance, and be done. Instead, null hypothesis testing is sometimes of little value (when testing models making quantitative predictions) and sometimes insufficient (for the practice application experiment). However, null hypothesis testing is ideal for supporting ordinal claims. Examples suggest that psychology might always have laws and theories making ordinal predictions, and in any case this is the situation most experimenters currently face.

Demonstrating Sufficient Evidence

A second major attack on null hypothesis testing begins with the assumption that the goal in science is to establish the degree to which a claim warrants belief. For example, in his criticism of null hypothesis testing. Rozeboom (1960) wrote, "The primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested" (p. 420).

It is useful to distinguish personal belief from warranted belief. *Personal belief* is what a person actually believes. *Warranted belief* is what a person should believe, given the available information. Thus, warranted belief is rational. If a personal belief was formed rationally, it would also be a warranted belief, but obviously personal beliefs are not always formed rationally.

Personal beliefs are obviously relevant to the pursuit of science, but they presumably should not be a part of formal science. For example, whether or not a manuscript is accepted for publication presumably should depend on the strength of evidence, the importance of the finding, and so on, not whether the conclusion matches the reviewer's personal beliefs. Therefore, the suggestions that beliefs should play a role in formal science presumably refer to warranted belief.

Degree of warranted belief in a claim would be quantified as the rationally estimated probability that a claim is correct. Therefore, the terms *warranted belief* and *probability correct* are interchangeable in the following discussions.

Null Hypothesis Testing: Not a Method of Supporting Beliefs

Null hypothesis testing is not a good method of estimating the probability that a claim is correct. First, according to the traditional interpretation, p is the probability of the observed results or larger given the null hypothesis, not the probability that the null hypothesis is correct. Therefore, null hypothesis testing does not determine the probability that a claim is correct (Bakan, 1966; Cohen, 1994; Oakes, 1986; Wilson, 1961).

Second, the overall probability that a claim is correct should take into account not only the evidence from a single experiment, but also evidence from previous experiments and the plausibility of the claim given other knowledge (Cox, 1958). Suppose one experimenter claims that eating Jell-O increases weight and another claims that eating Jell-O increases IQ, and both have evidence for their claim at $p = .04$. The former claim is more likely to be correct, because it is a priori plausible and the latter claim is not.

Third, null hypothesis testing has a "cliff" at .05. The value of p is compared to alpha, which is almost always set at .05. Success is then achieved when p is less than .05, creating a cliff: The value $p = .04$ is treated the same as $p = .001$, even though the two are very different; .06 is treated the same as .80, even though these two are also very different; and $p = .04$ is treated differently from $p = .06$, even though the two are nearly the same.

This cliff has often been criticized (e.g., Eysenck, 1960; Kempthorne, 1971; Morrison & Henkel, 1969; Oakes, 1986; Rozeboom, 1960). In fact, this cliff makes no sense from the standpoint of modifying beliefs. For example, $p = .001$ is a much better reason to modify a belief than $p = .04$, and $p = .04$ does not provide much more reason for modifying a belief than $p = .06$. Therefore, null hypothesis testing would not be the tool of choice if the goal in science was to determine the probability of a claim being correct.

Establishing Sufficient Evidence

If null hypothesis testing does not determine beliefs, what does it accomplish? It begins with the calculation of a test statistic, such as t or F. The size of

this statistic is a monotonically increasing function of the strength of evidence for an ordinal claim. For example, if two experimenters have identical experiments, the one with the higher value of t has the stronger evidence.

The value of p is defined as the probability of obtaining this value of the test statistic or larger if the null hypothesis is true. Converting this test statistic to a value of p creates a common measure of strength of evidence across statistical tests (Greenwald et al., 1996). Thus, $p = .04$ is equally strong evidence for a t test, F test, Mann-Whitney, or whatever statistical test is being used.

Finally, the obtained value of p is compared to a criterion alpha, which is conventionally set at .05 (for reasons to be explained later in this article). When p is less than .05, the experimenter has sufficient empirical evidence to support a claim. Thus, statistical testing functions to establish sufficient evidence to support a claim, with the criterion held constant across experiments.

Establishing a Corpus of Findings

Rozeboom (1960) suggested that scientists should not be making decisions about claims, they should be calculating and updating the probability of these claims. However, this does not seem practical. If there were only a handful of potential claims in any given area of psychology, it would be feasible to assign them probabilities, to be constantly updating the probabilities, and to expect experimenters to keep track of these ever-changing probabilities. In fact, just the number of claims in psychology is overwhelming. It would probably be impossible for human beings to keep track of the probability for each claim, especially if these probabilities were constantly changing. In any case, scientists do not assign probabilities to claims. Instead, scientists act like the goal of science is to collect a corpus of claims that are considered to be established (Giere, 1972).

It is useful to divide this corpus into two types of claims. The first type is general laws and theories, such as the law that frustration increases the tendency to aggression. These general laws and theories are essentially the knowledge of psychology. The second type of claim is experimental findings. For example, an experimental finding might be that, for these subjects in this experiment, frustration led to increased aggression, or a drug caused an improvement in IQ (as opposed to a difference between conditions being caused by just chance fluctuations). (The interpretation needed to allow null hypothesis testing to support this type of finding is presented in Frick, 1996.)

This corpus of findings provides the support for laws and theories. A scientist asserting some law or theory will often provide his or her own experimental finding as support. However, the scientist can cite other findings to support the law of theory, and there should not be any experimental findings that contradict the law or theory. Thus, using this corpus is both an opportunity and an obligation.

According to Cook and Campbell (1979), there are four steps to establishing validity. The first step, achieving statistical conclusion validity, is demonstrating that an effect is statistically significant. The second and third steps, achieving internal and construct validity, are verifying that there are no confounds. Experimental findings that pass these steps apparently are accepted as part of the corpus of psychology. Before publication, an experimental finding is examined by reviewers. If it does not meet these requirements it will not be published. Publication is then taken as certification by experts that the finding has passed these requirements. However, a finding's legitimacy (which is to say, its membership in the corpus of acceptable findings) can be later challenged by noting a confound or something inappropriate about the statistical test. If a finding is not published in a journal, it still can be accepted as a legitimate finding when these requirements are met.

This role of null hypothesis testing—as one of the criteria for a finding entering the corpus of psychology—explains the cliff (Chow, 1991). Cliffs are created by the need to make categorical decisions. For example, to be directly elected to the baseball hall of fame, a player must receive 75% of the votes of the baseball writers. A player either receives sufficient votes to be elected or does not not. With regard to assessing worthiness to enter the hall of fame, there is no cliff at 75%. However, with regard to eligibility to enter the hall of fame, there is a huge cliff at 75%. In null hypothesis testing, the categorical decision is between sufficient and insufficient evidence to support a finding. The line between the two is currently set at $p = .05$.

The Bayesian Alternative

The need for a categorical decision—about whether or not a claim should enter the corpus of psychology—does not rule out warranted belief as a criterion. In addition to other criteria, should the criterion for entering the corpus of psychology be warranted belief

or strength of evidence? If the answer is warranted belief, the next question is, What statistical test would be best for determining warranted belief?

A Bayesian statistical analysis (e.g., Edwards, Lindman, & Savage, 1963; Jeffreys, 1961; Oakes, 1986) takes into account the prior probabilities of the possible hypotheses, combines them with the available evidence, and yields (with suitable integration) the probability of the given claim being correct. One study found that, with any reasonable assumption about prior probabilities, a Bayesian analysis would outperform conventional statistical testing (Samaniego & Reneau, 1994). Therefore, if the goal was to estimate the probability that a claim was correct, a Bayesian analysis would be a very attractive alternative to null hypothesis testing.

However, there are several problems with a Bayesian analysis. First, the estimate of prior probabilities is subjective and somewhat arbitrary. As will be discussed, it is undesirable for the outcome of a statistical test to depend on subjective, arbitrary, or possibly biased choices by the experimenter. Of course, the current practice of forming beliefs is also subjective. Null hypothesis testing is used to evaluate the strength of evidence, and then combined subjectively with the results of other experiments and available theories, which would be presented along with the experimental finding. It is not clear which of these procedures is more subjective.

Second, it is not clear that a Bayesian analysis is more accurate than the current subjective practice. Oakes (1986) noted that constructing prior probabilities is not easy. It is somewhat contradictory to expect people to be able to intuitively construct good prior probabilities and yet require statistical testing to avoid the intuitive construction of final probabilities. Oakes also noted that there have been suggestions to adjust the prior probability so that the conclusion matches the user's expectations. This would undermine the rationale of a Bayesian analysis. Moreover, this suggestion implies that people can more accurately construct final beliefs than prior beliefs.

A third problem is the transient nature of the answer produced by the Bayesian analysis. The factor that would influence the prior probabilities would change across time, so the results of a Bayesian analysis would change across time. Thus, any Bayesian analysis would apply to only a short period of time. In contrast, the results of the null hypothesis test are an unchanging record of the amount of evidence supporting the finding. Obviously, it is more valuable to publish an unchanging statistic than one that quickly becomes outdated.

Thus, there are problems with using a Bayesian analysis. If beliefs should be the criterion, it is not clear that a Bayesian analysis would be better than the current practice.

Should Beliefs Be the Criterion?

The next question is, Should warranted belief be the criterion for a claim entering the corpus of psychology? For general laws and theories, the answer seems to be yes: Laws and theories presumably should be part of the corpus of psychology if and only if they are well-supported by all of the findings, not just the particular finding being presented by the experimenter.

However, there is a critical disadvantage to making warranted belief a criterion for findings entering the corpus of psychology (Oakes, 1986). Suppose the finding of one experiment would be consistent with current theories and expectations in psychology and the finding of a second experiment would contradict current theories and expectations. Suppose also that both findings achieve the same value of p, say $p = .02$. The first finding is more likely to be correct, which is to say warranted belief is higher for the first finding than the second. If warranted belief was the criterion for entering the corpus of psychology, the first finding would have an advantage over the second, and it is plausible that the second finding might not generate sufficient warranted belief to meet the criterion.

However, the second finding is the more valuable of the two. Findings that are consistent with current theories and expectations as a general rule do not advance a field of knowledge; the field is advanced by unexpected findings and findings that contradict current laws and theories. Therefore, if everything else is equal, unexpected findings are preferred over expected findings. The problem with using warranted belief as a criterion is that it gives a substantial advantage to expected findings over unexpected findings.

For example, consider Garcia's difficulties (described in Garcia, 1981) publishing his finding of learned taste aversion, which contradicted the dominant learning theory of the time. It is reasonable that psychologists of the time might not immediately believe Garcia's finding. However, it would have been unfortunate if his finding had not been published. More generally, science would not work well if the acceptance of a theory led to repression of findings

inconsistent with the theory and facilitation of findings consistent with the theory. Therefore, warranted belief would not be a good criterion for a finding entering the corpus of psychology.

Evaluation

Thus, psychology seems to work by having a corpus of claims. The criterion for belief in laws and theories probably should be warranted belief, but warranted belief would not be a good criterion for findings. The criterion for findings, which seems appropriate, is a lack of confounds and sufficient evidential support. Null hypothesis testing is then used to demonstrate sufficient evidential support, giving it an important and appropriate role in psychology.

Setting Alpha

In practice, the general rule is that alpha is set at .05 (Sterling, Rosenbaum, & Weinkam, 1995). However, many people believe that experimenters should be allowed to set their own value of alpha. For example, the APA *Publication Manual* (American Psychological Association, 1994) asks experimenters to report alpha, implying that they have a choice. The claim that experimenters should set their own value of alpha is not an attack on null hypothesis testing as described in textbooks, but it is an attack on null hypothesis testing as it is practiced in current psychology.

There are two reasons for adjusting alpha in statistical testing. The issue is whether they apply to the use of null hypothesis testing to demonstrate sufficient evidence for a finding.

Adjusting Alpha to Reflect Costs and Benefits

If a statistical test is being used to decide between actions, the decisions' costs (when wrong) and benefits (when correct) should be incorporated into alpha (Neyman & Pearson, 1933b; Oakes, 1986; Skipper, Guenther, & Nass, 1967). For example, if someone has a cold, it is relatively harmless to take vitamin C if it doesn't help and relatively valuable to take vitamin C if it does help. Thus, despite apparently uncertain evidence concerning the efficacy of vitamin C, many people take it for a cold.

However, suppose an experimenter wanted to publish a claim that taking vitamin C reduced the duration of a cold. Obviously, that experimenter should not be allowed to use a high value of alpha just because the action supported by that claim might be useful and cannot hurt. Instead, the claim should be backed by standard levels of evidence. Science is in the business of providing knowledge, not making decisions for people. The users of a finding can incorporate their own cost-benefit analysis into their decision.

The only action science takes is to publish a claim. Some claims are more important than others, which is to say they will attract more attention and use. Therefore, they have a higher benefit when correct. However, these claims also have a higher cost when wrong. Therefore, there is no obvious cost-benefit reason for why important claims should deserve a higher or lower value of alpha than any other claim.

Thus, it is important to consider costs and benefits when deciding between actions. However, there is little or no reason with respect to costs and benefits for adjusting alpha from one experiment to another.

Minimizing Error

A second reason for adjusting alpha is to minimize the total probability of error (Cohen, 1965; Neyman & Pearson, 1933a; Oakes, 1986; Winer, 1962). Suppose an experimenter is in the situation of choosing between two well-defined point hypotheses. For example, the experimenter might be trying to decide if a drug has (a) no effect on IQ or (b) raises IQ 6 points. In this situation, the experimenter can calculate the power of the experiment, which is defined as the likelihood of achieving statistical significance if the drug actually does increase IQ 6 points. The probability of making a Type II error (not rejecting the null hypothesis when the null hypothesis is incorrect) is equal to 1 − power. The total probability of making an error (the probability of making a Type I error plus the probability of making a Type II error) will be minimized by choosing a value of alpha that equalizes the two types of errors.

For example, suppose a very large number of subjects were being tested, such that, with alpha at .05, there would be 5% chance of making a Type I error and much less than a 1% chance of making a Type II error. The total probability of error would be between 5% and 6%. If alpha was lowered to .01, there would be only a 1% chance of making a Type I error and, assuming enough power, possibly still less than a 1% chance of making a Type II error, creating a total probability of error between 1% and 2%. Thus, adjusting alpha would lower the total probability of error. Similarly, if power was low, alpha could be raised to equalize the probability of Type I and Type II errors and lower the total probability of error.

However, experimenters usually are not in the sit-

uation of choosing between two well-defined point hypotheses (as noted by Neyman, 1950, p. 324). Instead, null hypothesis testing is used by experimenters to choose between the null hypothesis and some ordinal conclusion. Within this framework, there is no basis for determining the probability of a Type II error, hence no basis for adjusting alpha to lower total probability of error.

More important, the total probability of error is not the issue. First, when the null hypothesis is rejected, there is no chance of a Type II error. Second, null hypothesis testing is being used to demonstrate a sufficient amount of evidence. It would be inappropriate to choose a high value of alpha just to compensate for the fact that one was running an experiment with low power. It would similarly be inappropriate to ignore a finding at p less than .05 just because an experiment had high power.

Thus, when power can be known and when the goal is to reduce the total probability of error, it is useful to be able to adjust alpha. However, experimenters are not in this position—they are not choosing between two well-defined point hypotheses and they are not using null hypothesis testing to minimize the probability of error.

Objectivity

Thus, despite claims to the contrary (e.g., Labovitz, 1968), there is little reason for experimenters to choose different levels of alpha. On the other hand, allowing experimenters to choose their own alpha would inject a subjective element into statistical testing, which would be undesirable (Glass, McGaw, & Smith, 1981; Rozeboom, 1960). As Hick (1952) noted, the results of statistical testing should be determined only by the data, not the experimenter's opinion. Or, as Cox (1977) put it, two different experimenters should not reach different statistical conclusions given the same data.

Put another way, the question is, Should the experimenter decide what amount of evidence is sufficient for a finding to enter the corpus of psychology? Obviously not. In addition to being subjective and arbitrary, one might worry that experimenters would be biased in the selection of alpha for their own experiment. Furthermore, it would not work well if the experimenter could choose one value of alpha and other people citing the finding could choose a different level of alpha and come to a different conclusion. Therefore, to avoid the influence of the experimenter's opinions, judgments, and biases, it is desirable that experimenters not choose their own value of alpha. Instead, it is appropriate that alpha is set by the enterprise of psychology.

Summary

Null hypothesis testing can be easily criticized (a) assuming that it should be accomplishing something it does not, (b) assuming that experimenters should be accomplishing goals that do not require null hypothesis testing, or (c) not appreciating the situation in which null hypothesis testing is used. These criticisms usefully point out the limitations of null hypothesis testing. It is inappropriate for testing the quantitative predictions of models, it is insufficient for the practical application experiment, it is insufficient for determining warranted belief given all of the evidence, and an inflexible alpha is inappropriate when costs and benefits change or when the goal is to minimize overall error.

However, these restrictions do not rule out its use—null hypothesis testing is useful for demonstrating sufficient empirical evidence to support an ordinal claim. Because this is a common function in psychology, null hypothesis testing is appropriately used often. Experimenters commonly use a finding to test an ordinal law or the ordinal prediction of theory, findings are categorized as being acceptable or not to enter the corpus of claims in psychology, and sufficient evidence is an appropriate criterion for a claim being acceptable and warranted belief is not. The cost-benefit ratio does not change, the goal is not to minimize overall error, and subjective judgments by the experimenter are undesirable, so allowing the experimenter to select alpha is unneeded and inappropriate.

This article has not attempted to defend much of the philosophical justification typically associated with null hypothesis testing. Instead, it considered the practice of null hypothesis testing, justifying it only as an optimal procedure for assessing whether there is sufficient evidence in an experiment to support an ordinal finding.

References

American Psychological Association. (1994). *Publication manual.* Washington DC: Author.

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

Berkson, J. (1938). Some difficulties of interpretation en-

countered in the application of the chi-square test. *Journal of the American Statistical Association, 33,* 526–542.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin, 103,* 105–110.

Chow, S. L. (1991). Conceptual rigor versus practical impact. *Theory & Psychology, 1,* 337–360, 389–400.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics, 29,* 357–372.

Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statistics, 4,* 49–70.

Edwards, A. L. (1950). *Experimental design in psychological research.* New York: Rinehart.

Edwards, W., Lindman, H., & Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

Eysenck, H. J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review, 67,* 269–271.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5,* 75–98.

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

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition, 23,* 132–138.

Frick, R. W. (1996). Interpreting statistical testing: Not random sampling from a population. Submitted for publication.

Garcia, J. (1981). Tilting at the windmills of academe. *American Psychologist, 36,* 149–158.

Giere, R. N. (1972). The significance test controversy. *The British Journal for the Philosophy of Science, 23,* 170–181.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69,* 54–61.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p-values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175–183.

Harris, M. J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. *Theory & Psychology, 1,* 375–382.

Hick, W. E. (1952). A note on one-tailed and two-tailed tests. *Psychological Review, 59,* 316–318.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Jeffreys, H. (1961). *Theory of probability.* Oxford: Clarendon Press.

Johnstone, D. J. (1987). Tests of significance following R. A. Fisher. *The British Journal for the Philosophy of Science, 38,* 481–499.

Johnstone, D. J. (1988). Comments on Oakes on the foundations of statistical inference in the social and behavioral sciences: The market for statistical significance. *Psychological Reports, 63,* 319–331.

Kaiser, H. F. (1960). Directional statistical decision. *Psychological Review, 67,* 160–167.

Kalbfleisch, J. G., & Sprott, D. A. (1976). On test of significance. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 2, pp. 259–272). Dordrecht, Holland: Reidel.

Kempthorne, O. (1971). Probability, statistics, and the knowledge business. In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of statistical inferences* (pp. 470–490). Toronto: Holt, Rinehart, & Winston.

Labovitz, S. (1968). Criteria for selecting a significance level: A note on the sacredness of .05. *The American Sociologist, 3,* 200–222.

Levine, M. W., & Shefner, J. M. (1991). *Fundamentals of sensation and perception* (2nd ed.). Pacific Grove, CA: Brooks-Cole.

Loftus, G. R. (1994, August). Why psychology will never be a real science until we change the way we analyze data. Paper presented at the 102nd Annual convention of the American Psychological Association, Los Angeles, CA.

Loftus, G. R., & Masson, M. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490.

Morrison, D. E., & Henkel, R. E. (1969). Significance tests reconsidered. *The American Sociologist, 4,* 131–140.

Neyman, J. (1950). *First course in probability and statistics.* New York: Holt.

Neyman, J., & Pearson, E. S. (1933a). On the problem of the most efficient tests of statistical inference. *Biometrika, 20A,* 175–240, 263–294.

Neyman, J., & Pearson, E. S. (1933b). The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge philosphical society, 29,* 492–510.

Nunnally, J. C. (1960). The place of statistics in psychology. *Educational and Psychological Measurement, 20,* 641–650.

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

Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, & Computers, 28,* 12–22.

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

Samaniego, F. J., & Reneau, D. M. (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association, 89,* 947–957.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515–530.

Skipper, Jr., J. K., Guenther, A. L., & Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist, 2,* 16–18.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74,* 1–29.

Spjotvoll, E. (1977). Discussion of D. R. Cox's paper. *Scandinavian Journal of Statistics, 4,* 63–66.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49,* 108–112.

Tyler, R. W. (1931). What is statistical significance? *Educational Research Bulletin, 10,* 115–118, 142.

Weitzman, R. A. (1984). Seven treacherous pitfalls of statistics, illustrated. *Psychological Reports, 54,* 355–363.

Wilson, K. V. (1961). Subjectivist statistics for the current crisis. *Contemporary Psychology, 6,* 229–231.

Winer, B. J. (1962). *Statistical principles in experimental design.* New York: McGraw Hill.

Received June 6, 1995
Revision received May 20, 1996
Accepted June 5, 1996 ■

Correction to McGraw and Wong (1996)

The article "Forming Inferences About Some Intraclass Correlations Coefficients" by Kenneth O. McGraw and S. P. Wong (*Psychological Methods, 1996,* Vol. 1, No. 1, pp. 30–46) contained three errors. The intraclass correlation coefficient (ICC) and r values given in Table 6 (p. 39) of the article should be changed to $r = .714$ for each data set, ICC(C,1) = .714 for each data set, and ICC(A,1) = .720, .620, and .485 for the data in Columns 1, 2, and 3 of the table, respectively.

In Table 7 (p. 41), which is used to determine confidence intervals on population values of the ICC, the procedures for obtaining the confidence intervals on ICC(A,k) needs to be amended slightly. The definitions of F_* and F^* are said to be the same as for ICC(A,1); however, the degrees of freedom v need to be calculated using

$$c = \frac{\hat{p}}{n(1 - \hat{p})}$$

in place of a and

$$d = 1 + \frac{\hat{p}(n - 1)}{n(1 - \hat{p})}$$

in place of b.

On pages 44–46, references to Equations A3, A4, and so forth in the Appendix should be to Sections A3, A4, and so forth. We regret any inconvenience or confusion these errors may have caused.