General Article

An Alternative to Null-Hypothesis Significance Tests

Peter R. Killeen

Arizona State University

ABSTRACT—*The statistic* p_{rep} *estimates the probability of replicating an effect. It captures traditional publication criteria for signal-to-noise ratio, while avoiding parametric inference and the resulting Bayesian dilemma. In concert with effect size and replication intervals,* p_{rep} *provides all of the information now used in evaluating research, while avoiding many of the pitfalls of traditional statistical inference.*

Psychologists, who rightly pride themselves on their methodological expertise, have become increasingly embarrassed by "the survival of a flawed method" (Krueger, 2001) at the heart of their inferential procedures. Null-hypothesis significance tests (NHSTs) provide criteria for separating signal from noise in the majority of published research. They are based on inferred sampling distributions, given a hypothetical value for a parameter such as a population mean (μ) or difference of means between an experimental group (μ_E) and a control group (μ_C; e.g., H_0: $\mu_E - \mu_C = 0$). Analysis starts with a statistic on the obtained data, such as the difference in the sample means, D. D is a point on the line with probability mass of zero. It is necessary to relate that point to some interval in order to engage probability theory. Neyman and Pearson (1933) introduced critical intervals over which the probability of observing a statistic is less than a stipulated *significance level*, α (e.g., z scores between $[-\infty, -2]$ and between $[+2, +\infty]$ over which $\alpha < .05$). If a statistic falls within those intervals, it is deemed significantly different from that expected under the null hypothesis. Fisher (1959) preferred to calculate the probability of obtaining a statistic larger than $|D|$ over the interval $[|D|, \infty]$. This probability, $p(x \geq D|H_0)$, is called the p value of the statistic. Researchers typically hope to obtain a p value sufficiently small (viz. less than α) so that they can reject the null hypothesis.

Address correspondence to Peter Killeen, Department of Psychology, Arizona State University, Tempe, AZ 85287-1104; e-mail: killeen@asu.edu.

This is where problems arise. Fisher (1959), who introduced NHST, knew that "such a test of significance does not authorize us to make any statement about the hypothesis in question in terms of mathematical probability" (p. 35). This is because such statements concern $p(H_0|x \geq D)$, which does not generally equal $p(x \geq D|H_0)$. The confusion of one conditional for the other is analogous to the conversion fallacy in propositional logic. Bayes showed that $p(H|x \geq D) = p(x \geq D|H)p(H)/p(x \geq D)$. The unconditional probabilities are the *priors*, and are largely unknowable. Fisher (1959) allowed that $p(x \geq D|H_0)$ may "influence [the null's] acceptability" (p. 43). Unfortunately, absent priors, "P values can be highly misleading measures of the evidence provided by the data against the null hypothesis" (Berger & Selke, 1987, p. 112; also see Nickerson, 2000, p. 248). This constitutes a dilemma: On the one hand, "a test of significance contains no criterion for 'accepting' a hypothesis" (Fisher, 1959, p. 42), and on the other, we cannot safely reject a hypothesis without knowing the priors. Significance tests without priors are the "flaw in our method."

There have been numerous thoughtful reviews of this foundational issue (e.g., Nickerson, 2000), attempts to make the best of the situation (e.g., Trafimow, 2003), proposals for alternative statistics (e.g., Loftus, 1996), and defenses of significance tests and calls for their abolition alike (e.g., Harlow, Mulaik, & Steiger, 1997). When so many experts disagree on the solution, perhaps the problem itself is to blame. It was Fisher (1925) who focused the research community on parameter estimation "so convincingly that for the next 50 years or so almost all theoretical statisticians were completely parameter bound, paying little or no heed to inference about observables" (Geisser, 1992, p. 1). But it is rare for psychologists to need estimates of parameters; we are more typically interested in whether a causal relation exists between independent and dependent variables (but see Krantz, 1999; Steiger & Fouladi, 1997). Are women attracted more to men with symmetric faces than to men with asymmetric faces? Does variation in irrelevant dimensions of stimuli affect judgments on relevant dimensions? Does review of traumatic events facilitate recovery? Our unfortunate

historical commitment to significance tests forces us to rephrase these good questions in the negative, attempt to reject those nullities, and be left with nothing we can logically say about the questions—whether $p = .100$ or $p = .001$. This article provides an alternative, one that shifts the argument by offering "a solution to the question of replicability" (Krueger, 2001, p. 16).

PREDICTING REPLICABILITY

Consider an experiment in which the null hypothesis—no difference between experimental and control groups—can be rejected with a p value of .049. What is the probability that we can replicate this significance level? That depends on the state of nature. In this issue, as in most others, NHST requires us to take a stand on things that we cannot know. If the null is true, *ceteris paribus* we shall succeed—get a significant effect—5% of the time. If the null is false, replicability depends on the population effect size, δ. Power analysis varies the hypothetical discrepancy between the means of control and experimental populations, giving the probability of appropriately rejecting the null under those various assumptive states of nature. This awkward machinery is seldom invoked outside of grant proposals, whose review panels demand an n large enough to provide significant returns on funding.

Greenwald, Gonzalez, Guthrie, and Harris (1996) reviewed the NHST controversy and took the first clear steps toward a useful measure of replicability. They showed that p values predict the probability of getting significance in a replication attempt when the measured effect size, d', equals the population effect size, δ. This postulate, $\delta = d'$, complements NHST's $\delta = 0$, while making better use of the available data (i.e., the observed $d' > 0$). But replicating "significance" replicates the dilemma of significance tests: Data can speak to the probability of H_0 and the alternative, H_A, only after we have made a commitment to values of the priors. Abandoning the vain and unnecessary quest for definitive statements about parameters frees us to consider statistics that predict replicability in its broadest sense, while avoiding the Bayesian dilemma.

The Framework

Consider an experimental group and an independent control group whose sample means, M_E and M_C, differ by a score of D. The corresponding dimensionless measure of effect size d' (called d by Cohen, 1969; g by Hedges & Olkin, 1985; and d' in signal detectability theory) is

$$d' = \frac{M_E - M_C}{s_p}, \tag{1}$$

where s_p is the pooled within-group standard deviation. If the experimental and control populations are normal and the total sample size is greater than 20 ($n_E + n_C = n > 20$), the sampling distribution of d' is approximately normal (Hedges & Olkin, 1985; see the top panel of Fig. 1 and the appendix):

$$d' \sim N(\delta,\ \sigma_d). \tag{2}$$

σ_d is the standard error of the estimate of effect size, the square root of

$$\sigma_d{}^2 \approx \frac{n^2}{n_E n_C (n-4)}, \tag{3}$$

for $n > 4$. When $n_E = n_C$, Equation 3 reduces to $\sigma_d{}^2 \approx 4/(n-4)$.

Define *replication* as *an effect of the same sign as that found in the original experiment*. The probability of a replication attempt having an effect d'_2 greater than zero, given a population effect size of δ, is the area to the right of 0 in the sampling distribution centered at δ (middle panel of Fig. 1). Unfortunately, we do not know the value of the parameter δ and must therefore eliminate it.

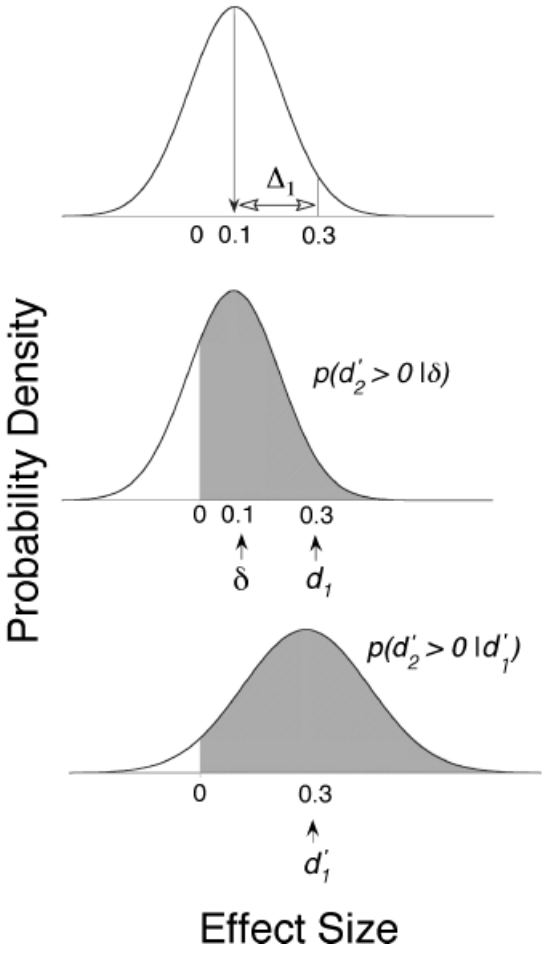

Fig. 1. Sampling distributions of effect size (d). The top panel shows a distribution for a population effect size of $\delta = 0.1$; the experiment yielded an effect size of 0.3, and thus had a sampling error $\Delta = d'_1 - \delta = 0.2$. The middle panel shows the probability of a replication as the area under the sampling distribution to the right of 0, given knowledge that $\delta = 0.1$. The bottom panel shows the posterior predictive density of effect size in replication. Absent knowledge of δ, the probability of replication is predicted as the area to the right of 0.

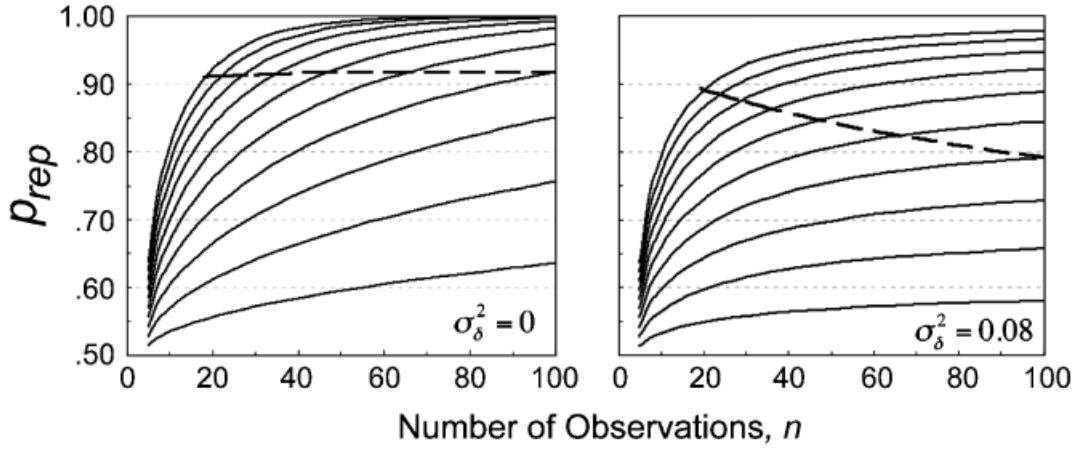

Fig. 2. Probability of replication (p_{rep}) as a function of the number of observations and measured effect size, d'_1. The functions in each panel show p_{rep} for values of d'_1 increasing in steps of 0.1, from 0.10 (lowest curve) to 1.0 (highest curve). The dashed lines show the combination of effect size and n necessary to reject a null hypothesis of no difference between the means of the experimental and control groups (i.e., $\mu_E - \mu_C = 0$) using a two-tailed t test with $\alpha = .05$. When realization variance, $\sigma_\delta{}^2$, is 0 (left panel), replicability functions asymptote at 1.0. For a one-tailed test, the dashed line drops to .88. When realization variance is 0.08 (right panel), the median for social psychological research, replicability functions asymptote below 1.0. As n approaches infinity, the t-test criterion falls to an asymptote of .5.

Eliminating δ

Define the sampling error, Δ, as $\Delta = d' - \delta$ (Fig. 1, top panel). For the original experiment, this equation may be rewritten as $\delta = d'_1 - \Delta_1$. Replication requires that if d'_1 is greater than 0, then d'_2 is also greater than 0, that is, that $d'_2 = \delta + \Delta_2 > 0$. Substitute $d'_1 - \Delta_1$ in place of δ in this equation. Replication thus requires that $d'_2 = d'_1 - \Delta_1 + \Delta_2 > 0$. The expectation of each sampling error is 0 with variance $\sigma_d{}^2$. For independent replications, the variances add, so that $d'_2 \sim N(d'_1, \sigma_{d_R})$, with $\sigma_{d_R} = \sqrt{2}\sigma_{d_R}$. The probability of replication, p_{rep}, is the area of the distribution for which d' is greater than 0, shaded in the bottom panel of Figure 1:

$$p_{rep} = \int_0^\infty n(d'_1,\ \sigma_{d_R}). \tag{4}$$

Slide the distribution to the left by the distance d'_1 to see that Equation 4 describes the same area as

$$p_{rep} = \int_{-d'_1}^\infty n(0,\ \sigma_{d_R}) = \int_{-\infty}^{d'_1} n(0,\ \sigma_{d_R}). \tag{5}$$

It is easiest to calculate p_{rep} from the right integral in Equations 5, by consulting a normal probability table for the cumulative probability up to

$$z = \frac{d'_1}{\sigma_{d_R}}. \tag{6}$$

Example

Suppose an experiment with $n_E = n_C = 12$ yields a difference between experimental and control groups of 5.0 with $s_p = 10.0$. This gives an effect of $d'_1 = 0.5$ (Equation 1) with a variance of $\sigma_{d_1}{}^2 \approx 4/(24 - 4) = 0.20$ (Equation 3), and a replication variance of $\sigma_{d_R}{}^2 = 2 \cdot \sigma_{d_1}{}^2 \approx 0.40$. From this, it follows that $z = 0.5/\sqrt{0.40} = 0.79$ (Equation 6). A table of the normal distribution assigns a p_{rep} of .785.[1]

As the hypothetical number of observations in the replicate approaches infinity, the sampling variance of the replication goes to zero, and p_{rep} is the positive area of $N(d'_1, \sigma_{d_1})$. This is the sampling distribution of a standard power analysis at the maximum likelihood value for δ, and establishes an upper bound for replicability. It is unlikely, however, that the next investigator will have sufficient resources or interest to approach that upper bound. By default, then, p_{rep} is defined for equipotent replications, ones that employ the same number of subjects as the original experiment and experience similar levels of sampling error. The probability of replication may be calculated under other scenarios (as shown later), but for purposes of qualifying the data in hand, equipotency, which doubles the sampling variance, is assumed.

The left panel of Figure 2 shows the probability of replicating the results of an experiment whose measured effect size is $d'_1 =$ 0.1 (bottom curve), 0.2, . . . , 1.0, as a function of the number of observations in the original study. These results permit a comparison with traditional measures of significance. The dashed line connects the effect sizes necessary to reject the null under a two-tailed t test, with probability of a Type I error, α, less than .05. Satisfying this criterion is tantamount to establishing a p_{rep} of approximately .917.

Parametric Variance

The calculations presented thus far assume that the variance contributed by contextual variables in the replicate is negligible

[1]Excel® spreadsheets with relevant calculations are available from http://www.asu.edu/clas/psych/research/sqab and from http://www.latrobe.edu.au/psy/esci/.

compared with the sampling error of d. This is the classic fixed-effects model of science. But every experiment is a sample from a population of possible experiments on the topic, and each of those, with its own differences in detail, has its own subspecies of effect size, δ_i. This is true a fortiori for correlational studies involving different instruments or moderators (Mosteller & Colditz, 1996). The population of effect sizes adds a *realization variance*, $\sigma_\delta{}^2$, to the sampling distributions of the original and the replicate (Raudenbush, 1994; Rubin, 1981; van den Noortgate & Onghena, 2003), so that the standard error of effect size in replication becomes

$$\sigma_{d_R} = \sqrt{2(\sigma_{d_1}{}^2 + \sigma_{d_i}{}^2)}. \quad (7)$$

In a recent meta-meta-analysis of more than 25,000 social science studies, Richard, Bond, and Stokes-Zoota (2003) reported a mean *within-literature variance* of $\sigma_\delta{}^2 = 0.092$ (median = 0.08), corrected for sampling variance (Hedges & Vevea, 1998). The statistic $\sigma_\delta{}^2$ places an upper limit on the probability of replication, one felt most severely by studies with small effect sizes. This is shown graphically in the right panel of Figure 2. The probability of replication no longer asymptotes at 1.0, but rather at $p_{\mathrm{rep(max)}} = \int_{-\infty}^{d'_1} n(0,\sqrt{2}\sigma_\delta)$. At $n = 100$, the functions shown in the right panel of Figure 2 are no more than 5 points below their asymptotes. Given a representative $\sigma_\delta{}^2$ of 0.08, for no value of n will a measured effect size of d' less than 0.52 attain a p_{rep} greater than .90; but this standard comes within reach of a sample size of 40 for a d' of 0.8.

Reliance on standard hypothesis-testing techniques that ignore realization variance may be one of the causes for the dismayingly common failures of replication. The standard t test will judge an effect of any size significant at a sufficiently large n, even though the odds for replication may be very close to chance. Figure 2 provides understanding, if no consolation, to investigators who have failed to replicate published findings of high significance but low effect size. The odds were never very much in their favor. Setting a replicability criterion for publication that includes an estimate of realization variance would filter the correlational background noise noted by Meehl (1997) and others.

Claiming replicability for an effect that would merely be of the same sign may seem too liberal, when the prior probability of that is 1/2, but traditional null-hypothesis tests are themselves at best merely directional. The proper metric of effect size is d or r, not p or p_{rep}. In the present analysis, replicability qualifies effect, not effect size: A d'_2 of 2.0 constitutes a failure to replicate an effect size (d'_1) of 0.3, but is a strong replication of the effect. Requiring a result to have a p_{rep} of .9 exacts a standard comparable to (Fig. 2, left panel) or exceeding (right panel) the standard of traditional significance tests.

Does p_{rep} really predict the probability of replication? In a meta-analysis of 37 studies of the psychophysiology of aggression, including unpublished nonsignificant data sets, Lorber (2004) found that 70% showed a negative relation between heart rate and aggressive behavior patterns. The median value of p_{rep} over those studies was .71 (.69 assuming $\sigma_\delta{}^2 = 0.08$). In a meta-analysis of 37 studies of the effectiveness of massage therapy, Moyer, Rounds, and Hannum (2004) found that 83% reported positive effects on various dependent variables; including an estimate of publication bias against negative results reduced this value to 74%. The median value of p_{rep} over those studies was .75 (.73 assuming $\sigma_\delta{}^2 = 0.08$). In a meta-analysis of 45 studies of transformational leadership, Eagly, Johannesen-Schmidt, and van Engen (2003) found that 82% showed an advantage for women, and argued against attenuation by publication bias. The median value of p_{rep} over these studies was .79 (dropping to .68 for $\sigma_\delta{}^2 = 0.08$ because of the generally small effect sizes). Averaging values of p_{rep} and counting the proportion of positive results are both inefficient ways of aggregating and evaluating data (Cooper & Hedges, 1994), but such analyses provide face validity for p_{rep}, which is intended primarily as a measure of the robustness of studies taken singly.

Generalizations

Whenever an effect size can be calculated (see Rosenthal, 1994, for conversions among indices; Cortina & Nouri, 2000, for analysis of variance designs; Grissom & Kim, 2001, for caveats), so also can p_{rep}. Randomization tests, described in the appendix, facilitate computation of p_{rep} for complex designs or situations in which assumptions of normality are untenable. Calculation of the n required for a desired p_{rep} is straightforward. For a presumptive effect size of δ and realization variance of $\sigma_\delta{}^2$, calculate the z score corresponding to p_{rep}, and employ an $n = n_E + n_C$ no fewer than

$$n = \frac{8z^2}{\delta^2 - 2\sigma_\delta{}^2 z^2} + 4. \quad (8)$$

Negative results indicate that the desired p_{rep} is unobtainable for that $\sigma_\delta{}^2$. For example, for $\delta = 0.8$, $\sigma_\delta{}^2 = 0.08$, and a desired $p_{\mathrm{rep}} = .9$, $z(.9)^2 = 1.64$, and the minimum n is 40.

Stronger claims than replication of a positive effect are sometimes warranted. An investigator may wish to claim that a new drug is more effective than a standard. The replicability of the data supporting that claim may be calculated by integrating Equation 4 not from 0, but from d_s, the effect size of the standard bearer. Editors may prefer to call a result replicable only if it accounts for, say, at least 1% of the variance in the data, for which d' must be greater than 0.04. They may also require that it pass the Aikaike criterion for adding a parameter (distinct means for experimental and control groups; Burnham & Anderson, 2002), for which r^2 must be greater than $1 - e^{-2/n}$. Together, these constraints define a lower limit for "replicable" at $p_{\mathrm{rep}} \approx 55$. However these minima are set, a fair assessment of σ_δ is necessary for p_{rep} to give investigators a fair assessment of replicability.

The replicability of differences among experimental conditions is calculated the same way as that between experimental and control conditions. Multiple comparisons are made by the conjunction or disjunction of p_{rep}: If treatments A and B are independent, each with p_{rep} of .80, the probability of replicating both effects is .64, and the probability of replicating at least one is .87. The probability of n independent attempts to replicate an experiment all succeeding is ${p_{rep}}^n$.

As is the case for all statistics, there is sampling variability associated with d', so that any particular value of p_{rep} may be more or less representative of the values found by other studies executed under similar conditions. It is an estimate. Replication intervals (RIs) aid interpretation by reflecting p_{rep} onto the measurement axis. Their calculation is the same as for confidence intervals (CIs), but with variance doubled. RIs can be used as equivalence tests for evaluating point predictions. The standard error of estimate conveniently captures 52% of future replications (Cumming, Williams, & Fidler, 2004). This familiar error bar can therefore be interpreted as an approximate 50% RI. In the example given earlier, for $\sigma_\delta = 0$, the 50% RI for D is approximately $5 \pm \sqrt{2(10^2/24)} \approx [2.1,\ 7.9]$.

WHY SWITCH?

Sampling distributions for replicates involve two sources of variance, leading to a root-2 increase in the standard error over that used to calculate significance. Why incur that cost? Both p and p_{rep} are functions of effect size and n, and so convey similar information: The top panel in Figure 3 shows p as the area in the right tail of the sampling distribution of d'_1, given the null, and p_{rep} as the area in the right tail of the prospective sampling distribution of d'_2, given d'_1. As d'_1 or n varies, p_{rep} and p change in complement.

Recapturing a familiar index of merit is reassuring, as are the familiar calculations involved; but these analyses are not equivalent. Consider the following contrasts:

Intuitive Sense

What is the difference between p values of .05 and .01, or between p values of .01 and .001? If you follow Neyman-Pearson and have set α to be .05, you must answer, "Nothing" (Meehl, 1978). If you follow Fisher, you can say, "The probability of finding a statistic more extreme than this under the null is p." Now compare those p values, and the oblique responses they support, with their corresponding values of p_{rep} shown in the bottom panel of Figure 3. These steps in p values take us from p_{rep} of .88 to .95 to .99—increments that are clear, interpretable, and manifestly important to a practicing scientist.

Logical Authority

Under NHST, one can never accept a hypothesis, and is often left in the triple-negative no-man's land of failure to reject the null. The p_{rep} statistic provides a graded measure of replicability that authorizes positive statements about results: "This effect will replicate $100(p_{rep})$% of the time" conveys useful information, whatever the value of p_{rep}.

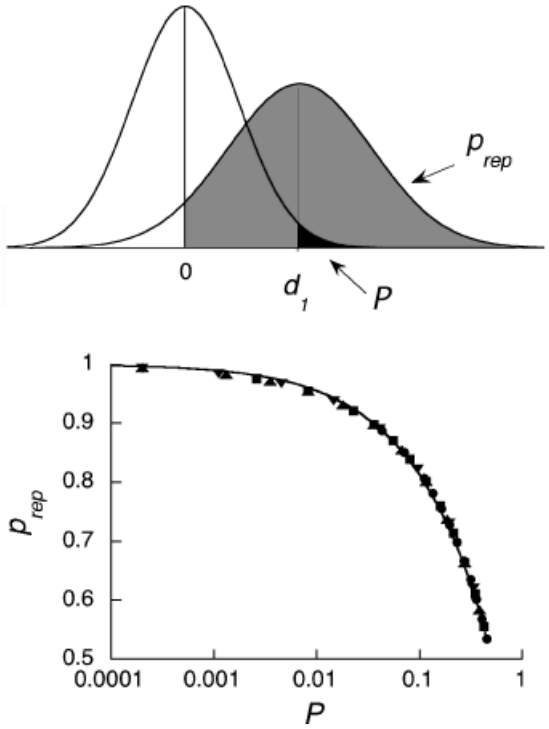

Fig. 3. Complementarity of p_{rep} and p. The top panel shows sampling distributions for d'_1 given the null (left) and for d'_2 given d_1 (right). The small black area gives the probability of finding a statistic more extreme than d_1 if the null were true. The large shaded area gives the probability of finding supportive evidence in an equipotent replication. In the bottom panel, p_{rep} is plotted against the p values calculated for the normal distribution under the null hypothesis with $d = 0.1, 0.2, \ldots, 1.0$, and n ranging from 10 to 80; p_{rep} is calculated from Equations 3, 5, and 6. The function is described in the appendix.

Real Power

Traditionally, replication has been viewed as a second successful attainment of a significant effect. The probability of getting a significant effect in a replicate is found by integrating Equation 4 from a lower limit given by the critical value $d^* = \sigma_{d_R} t_{\alpha,\nu_2-2}$. This calculation does not require that the original study achieved significance. Such analyses may help bridge to the new perspective; but once p_{rep} is determined, calculation of traditional significance is a step backward. The curves in Figure 2 predict the replicability of an effect given known results, not the probability of a statistic given the value of a parameter whose value is not given.

Elimination of Errors

Significance level is defined as the probability of rejecting the null when it is true (a Type I error of probability α); power is defined as the probability of rejecting the null when it is false,

and not doing so is a Type II error. False premises lead to conclusions that may be logically consistent but empirically invalid, a Type III error. Calculations of p are contingent on the null being true. Because the null is almost always false (Cohen, 1994), investigators who imply that manipulations were effective on the basis of a p less than α are prone to Type III errors. Because p_{rep} is not conditional on the truth value of the null, it avoids all three types of error.

One might, of course, be misled by a value of p_{rep} that itself cannot be replicated. This can be caused by

- sampling error: d_1 may deviate substantially from δ (RIs help interpret this risk.)
- failure to include an estimate of $\sigma_\delta{}^2$ in the replication variance
- publication bias against small or negative effects
- the presence of confounds, biased data selection, and other missteps that plague all mapping of particular results to general claims

Because of these uncertainties, p_{rep} is only an estimate of the proportion of replication attempts that will be successful. It measures the robustness of a demonstration; its accuracy in predicting the proportion of positive replications depends on the factors just listed.

Greater Confidence

The American Psychological Association (Wilkinson & the Task Force on Statistical Inference, 1999) has called for the increased use of CIs. Unfortunately, few researchers know how to interpret them, and fewer still know where to put them (Cumming & Finch, 2001; Cumming et al., 2004; Estes, 1997; Smithson, 2003; Thompson, 2002). CIs are often drawn centered over the sample statistic, as though it were the parameter; when a CI does not subsume 0, it is often concluded that the null may be rejected. The first practice is misleading, and the second wrong. CIs are derived from sampling distributions of M around a hypostatized μ: $|\mu - M|$ will be less than the CI $100p\%$ of the time. But as difference scores, CIs have lost their location. Situating them requires an implicit commitment to parameters—either to $\mu = 0$ for NHST or to $\mu = M$ for the typical position of CIs flanking the statistic. Such a commitment, absent priors, runs afoul of the Bayesian dilemma. In contrast, RIs can be validly centered on the statistic to which they refer, and the replication level may be correctly interpreted as the probability that the statistics of future equipotent replications will fall within the interval.

Decision Readiness

Significance tests are said to provide decision criteria essential to science. But it is a poor decision theory that takes no account of prior information and no account of expected values, and in the end lets us decide only whether or not to reject a statistic as improbable under the null. As a graduated measure, p_{rep} provides a basis for a richer approach to decision making than the Neyman-Pearson strategy, currently the mode in psychology. Decision makers may compute expected value, $E(v)$, by multiplying p_{rep} or its complement by the values they assign outcomes. Let $v^+(d')$ be the value of positive action for an effect size d', including potential costs for small or contrary effects. Then $E(v^+) = \int_{-\infty}^{+\infty} v^+(x)n\left(x; d'_1, \sigma_R\right)$. Comparison with an analogous calculation for $E(v^-)$ will inform the decision.

Congeniality With Bayes

Probability theory provides a unique basis for the logic of science (Cox, 1961), and Bayes' theorem provides the machinery to make science cumulative (Jaynes & Bretthorst, 2003; see the appendix). Falsification of the null cannot contribute to the cumulation of knowledge (Stove, 1982); the use of Bayes to reduce $\sigma_{d_R}{}^2$ can. NHST stipulates an arbitrary mean for the test statistic a priori (0) and a variance a posteriori $\left(s_p{}^2/n\right)$. The statistic p_{rep} uses both moments of the observed data in a coherent fashion to predict the most likely posterior distribution of the replicate statistic. Information from replicates may be pooled to reduce $\sigma_d{}^2$ (Louis & Zelterman, 1994; Miller & Pollack, 1994). Systematic explorations of phenomena identify predictors or moderators that reduce $\sigma_\delta{}^2$. The information contributed by an experiment, and thus its contribution to knowledge, is a direct function of this reduction in $\sigma_{d_R}{}^2$.

Improved Communication

The classic definition of replicability can cause harmful confusion when weak but supportive results must be categorized as a "failure to replicate [at $p < .05$]" (Rossi, 1997). Consider an experiment involving memory for deep versus superficial encoding of target words. This experiment, conducted in an undergraduate methods class, yielded a highly significant effect for the pooled data of 124 students, $t(122) = 5.46$ (Parkinson, 2004). We can "power down" the effect estimated from the pooled data to predict the probability that each of the seven sections in which these data were collected would replicate this classic effect. All of the test materials and instructions were identical, so $\sigma_\delta{}^2$ was approximately 0. The effect size from the pooled data, d', was 0.49. Individual class sections, averaging ns of 18, contributed the majority of variability to the replicate sampling distribution, whose variance is the sum of sampling variances for $n = 124$ ("original") and again for $n = 18$ (replicates). Replacing σ_{d_R} in Equation 4 with the root of this sum predicts a replicability of .81: Approximately six of the seven sections should get a positive effect. It happens that all seven did, although for one the effect size was a mere 0.06. Unfortunately, the instructor had to tell four of the seven sections that they had, by contemporary standards, failed to replicate a very

reliable result, as their *p*s were greater than .05. It was a good opportunity to discuss sampling error. It was not a good opportunity to discuss careers in psychology.

"How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service!" (Darwin, 1994, p. 269). Significance tests can never be *for*: "Never use the unfortunate expression 'accept the null hypothesis'" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599). And without priors, there are no secure grounds for being *against*—rejecting— the null. It follows that if our observations are to be of any service, it will not be because we have used significance tests. All this may be hard news for small-effects research, in which significance attends any hypothesis given enough *n*, whether or not the results are replicable. But editors may lower the hurdle for potentially important research that comes with so precise a warning label as p_{rep}. When replicability becomes the criterion, researchers can gauge the risks they face in pursuing a line of study: An assistant professor may choose paradigms in which p_{rep} is typically greater than .8, whereas a tenured risk taker may hope to reduce $\sigma_\delta{}^2$ in a line of research having p_{rep}s around .6. When replicability becomes the criterion, *significance,* shorn of its statistical duty, can once again become a synonym for the importance of a result, not for its improbability.

Acknowledgments—Colleagues whose comments have improved this article include Sandy Braver, Darlene Crone-Todd, James Cutting, Randy Grace, Tony Greenwald, Geoff Loftus, Armando Machado, Roger Milsap, Ray Nickerson, Morris Okun, Clark Presson, Anon Reviewer, Matt Sitomer, and François Tonneau. In particular, I thank Geoff Cumming, whose careful readings saved me from more than one error. The concept was presented at a meeting of the Society of Experimental Psychologists, March 2004, Cornell University. The research was supported by National Science Foundation Grant IBN 0236821 and National Institute of Mental Health Grant 1R01MH066860.

REFERENCES

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

Bruce, P. (2003). Resampling stats in Excel [Computer software]. Retrieved February 1, 2005, from http://www.resample.com

Burnham, K.P., & Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer-Verlag.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

Cooper, H., & Hedges, L.V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cortina, J.M., & Nouri, H. (2000). *Effect size for ANOVA designs*. Thousand Oaks, CA: Sage.

Cox, R.T. (1961). *The algebra of probable inference*. Baltimore: Johns Hopkins University Press.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532–575.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, *3*, 299–311.

Darwin, C. (1994). *The correspondence of Charles Darwin* (Vol. 9; F. Burkhardt, J. Browne, D.M. Porter, & M. Richmond, Eds.). Cambridge, England: Cambridge University Press.

Eagly, A.H., Johannesen-Schmidt, M.C., & van Engen, M.L. (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing men and women. *Psychological Bulletin*, *129*, 569–591.

Estes, W.K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review*, *4*, 330–341.

Fisher, R.A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700–725.

Fisher, R.A. (1959). *Statistical methods and scientific inference* (2nd ed.). New York: Hafner Publishing.

Geisser, S. (1992). Introduction to Fisher (1922): On the mathematical foundations of theoretical statistics. In S. Kotz & N.L. Johnson (Eds.), *Breakthroughs in statistics* (Vol. 1, pp. 1–10). New York: Springer-Verlag.

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

Grissom, R.J., & Kim, J.J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, *6*, 135–146.

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect sizes and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hedges, L.V., & Vevea, J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Jaynes, E.T., & Bretthorst, G.L. (2003). *Probability theory: The logic of science*. Cambridge, England: Cambridge University Press.

Krantz, D.H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *44*, 1372–1381.

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

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Lorber, M.F. (2004). Psychophysiology of aggression, psychopathy, and conduct problems: A meta-analysis. *Psychological Bulletin*, *130*, 531–552.

Louis, T.A., & Zelterman, D. (1994). Bayesian approaches to research synthesis. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 411–422). New York: Russell Sage Foundation.

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

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

Miller, N., & Pollock, V.E. (1994). Meta-analytic synthesis for theory development. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 457–484). New York: Russell Sage Foundation.

Mosteller, F., & Colditz, G.A. (1996). Understanding research synthesis (meta-analysis). *Annual Review of Public Health*, *17*, 1–23.

Moyer, C.A., Rounds, J., & Hannum, J.W. (2004). A meta-analysis of massage therapy research. *Psychological Bulletin*, *130*, 3–18.

Neyman, J., & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, *231*, 289–337.

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

Parkinson, S.R. (2004). [Levels of processing experiments in a methods class]. Unpublished raw data.

Raudenbush, S.W. (1994). Random effects models. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Richard, F.D., Bond, C.F., Jr., & Stokes-Zoota, J.J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosenthal, R., & Rubin, D.B. (2003). $r_{equivalent}$: A simple effect size indicator. *Psychological Methods*, *8*, 492–496.

Rossi, J.S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 175–197). Mahwah, NJ: Erlbaum.

Rubin, D.B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, *6*, 377–400.

Smithson, M. (2003). *Confidence intervals*. Thousand Oaks, CA: Sage.

Steiger, J.H., & Fouladi, R.T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.

Stove, D.C. (1982). *Popper and after: Four modern irrationalists*. New York: Pergamon Press (Available from Krishna Kunchithapadam, http://www.geocities.com/ResearchTriangle/Facility/4118/dcs/popper)

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25–32.

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

van den Noortgate, W., & Onghena, P. (2003). Estimating the mean effect size in meta-analysis: Bias, precision, and mean squared error of different weighting methods. *Behavior Research Methods, Instruments, & Computers*, *35*, 504–511.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

(Received 5/5/04; Revision accepted 7/30/04)

APPENDIX

This back room contains equations, details, and generalizations.

Effect Size

The denominator of effect size given by Equation 1 is the pooled variance, calculated as

$$s_p^2 = \frac{s_C^2(n_C - 1) + s_E^2(n_E - 1)}{n - 2}.$$

Hedges (1981) showed that an unbiased estimate of δ is

$$d \approx d'[1 - 3/(4n - 9)].$$

The adjustment is small, however, and with suitable adjustments in σ_d, d' suffices.

Negative effects generate p_{rep}s less than .5, indicating the unlikelihood of positive effects in replication. For consistency, if d' is less than 0, use $|d'|$ and report the result as the replicability of a negative effect. Useful conversions are $d' = 2r(1 - r^2)^{-1/2}$ (Rosenthal, 1994) and $d' = t[1/n_E + 1/n_C]^{1/2}$ for the simple two-independent-group case and $d' = t_r[(1 - r)/n_E + (1 - r)/n_C]^{1/2}$ for a repeated measures t, where r is the correlation between the measures (Cortina & Nouri, 2000).

The *asymptotic variance* of effect size (Hedges, 1981) is

$$\sigma_d^2 = \frac{n}{n_E n_C} + \frac{d^2}{2n}.$$

Equation 3 in the text is optimized for the use of d', however, and delivers accurate values of p_{rep} for $-1 \le d' \le 1$.

Variance of Replicates

The desired variance of replicates, $\sigma_{d_R}^2$, equals the expectation $E\left[(d_2 - d_1)^2\right]$. This may be expanded (Estes, 1997) as

$$\begin{aligned} E\left[(d_2 - d_1)^2\right] &= E\left[((d_2 - \delta) - (d_1 - \delta))^2\right] \\ &= E\left[(d_2 - \delta)^2 + (d_1 - \delta)^2\right] \\ &\quad - 2E[(d_2 - \delta)(d_1 - \delta)] \end{aligned}$$

The quantities $E\left[(d_2 - \delta)^2\right]$ and $E\left[(d_1 - \delta)^2\right]$ are the variances of d_2 and d_1, each equal to σ_d^2. For independent replications, the expectation of the cross product $E[(d_2 - \delta)(d_1 - \delta)]$ is 0.

Therefore, $\sigma_{d_R}^2 = E\left[(d_2 - d_1)^2\right] = \sigma_d^2 + \sigma_d^2$. It follows that the standard error of effect size of equipotent replications is $\sigma_{d_R} = \sqrt{2}\sigma_d$.

When $n_E = n_C > 2$,

$$\sigma_{d_R}{}^2 \approx \frac{8}{n-4} + 2\sigma_\delta{}^2$$

When the sizes of the original and replicate samples vary, replication variance should be based on

$$\sigma_{d_R}{}^2 = \sigma_{d_1',n_1}{}^2 + \sigma_{d_1',n_2}{}^2 + 2\sigma_\delta{}^2.$$

p_{rep} as a Function of p

We may approximate the normal distribution by the logistic and solve for p_{rep} as a function of p. This suggests the following equation:

$$p_{rep} \approx \left[1 + \left(\frac{p}{1-p}\right)^{2/3}\right]^{-1}.$$

The parenthetical converts a p value into a probability ratio appropriate for the logistic inverse. For two-tailed comparisons, halve p. Users of Excel can simply evaluate p_{rep} = NORMSDIST(NORMSINV(1 − p)/SQRT(2)) (G. Cumming, personal communication, October 24, 2004). This estimate is complementary to Rosenthal and Rubin's (2003) estimate of effect size directly from p and n.

Randomization Method

Randomization methods avoid assumptions of normality, are useful for small-n experiments, and are robust against heteroscedasticity. To employ them:

- Bootstrap populations for the experimental and control samples independently, generating subsamples of half the size of the original samples, using software such as Resampling Stats© (Bruce, 2003). This half-sizing provides the $\sqrt{2}$ increase in the standard deviation intrinsic to calculation of p_{rep}.
- Generate an empirical sampling distribution of the difference of the means of the subsamples, or of the mean of the differences for a matched-sample design.
- The proportion of the means that are positive gives p_{rep}.

This robust approach does not take into account $\sigma_\delta{}^2$, and so is accurate only for exact replications.

A Cumulative Science

Falsification of the null, even when possible, provides no machinery for the cumulation of knowledge. Reduction of σ_{d_R} does. Information is the reduction of entropy, which can be measured as the Fisher information content of the distribution of effect sizes. The difference of the entropies before and after an experiment, $I = log_2(\sigma_{before}/\sigma_{after})$, measures its incremental contribution of information. The discovery of better theoretical structures, predictors, or moderators that convert within-group variance to between-group variance permits large reductions in $\sigma_\delta{}^2$, and thus σ_{d_R}; smaller reductions are effected by cumulative increases in n.