

Killeen's (2005) p_{rep} Coefficient: Logical and Mathematical Problems

Michael Maraun and Stephanie Gabriel
Simon Fraser University

In his article, "An Alternative to Null-Hypothesis Significance Tests," Killeen (2005) urged the discipline to abandon the practice of p_{obs} -based null hypothesis testing and to quantify the signal-to-noise characteristics of experimental outcomes with replication probabilities. He described the coefficient that he invented, p_{rep} , as the probability of obtaining "an effect of the same sign as that found in an original experiment" (Killeen, 2005, p. 346). The journal *Psychological Science* quickly came to encourage researchers to employ p_{rep} , rather than p_{obs} , in the reporting of their experimental findings. In the current article, we (a) establish that Killeen's derivation of p_{rep} contains an error, the result of which is that p_{rep} is not, in fact, the probability that Killeen set out to derive; (b) establish that p_{rep} is not a replication probability of any kind but, rather, is a quasi-power coefficient; and (c) suggest that Killeen has mischaracterized both the relationship between replication probabilities and statistical inference, and the kinds of claims that are licensed by knowledge of the value assumed by the replication probability that he attempted to derive.

In his article, "An Alternative to Null-Hypothesis Significance Tests," Peter Killeen (2005) suggested that at least some of the defects inherent to the practice of null hypothesis significance testing, the brand of inference that has come to dominate work within the social and behavioral sciences (see, e.g., Berger & Selke, 1987; Krueger, 2001; Loftus, 1996; Nickerson, 2000), could be overcome by a reorientation of the inferential problem. Following the lead of Greenwald, Gonzalez, Harris, and Guthrie (1996), Killeen suggested that the researcher shift his or her focus from parameters to observables and quantify the signal-to-noise properties of experimental outcomes as a coefficient of replicability. The product of this conceptual shift in focus was a new coefficient, p_{rep} , which Killeen (2005) described as the probability of obtaining "an effect of the same sign as that found in an original experiment" (p. 346). According to Killeen (2005), " p_{rep} . . . 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" (p. 345). Despite the fact that two of the commentators on Killeen's article, Doros and Geier (2005) and Macdonald (2005), presented analyses that appeared to reveal errors in Killeen's derivation of p_{rep} , the journal *Psychological Science* nevertheless adopted the stance that researchers "are encouraged to use p_{rep} rather than p values" (Cutting, 2005) in their submissions to the journal.

Although we agree that an approach to inference that is dominated by null hypothesis significance testing is badly in need of reform,

Killeen's (2005) coefficient p_{rep} is most certainly not the solution. We will, herein: (a) establish that Killeen's derivation of p_{rep} contains an error the result of which is that p_{rep} is not, in fact, the probability that Killeen set out to derive; (b) establish that p_{rep} is not a replication probability of any kind, but is, rather, a quasi-power coefficient, very much the same kind of coefficient that researchers currently compute in their power analyses; (c) suggest that Killeen has mischaracterized both the relationship between replication probabilities and statistical inference and the kinds of claims that are licensed by the possession of knowledge of the value assumed by the replication probability that he attempted to derive. Before turning to the task of elucidating these charges, we review the basis for Killeen's very reasonable dissent over the inferential machinery that is indigenous to the social and behavioral sciences and provide a point-for-point account of his derivation of p_{rep} .

Killeen's (2005) Motivation for Inventing p_{rep}

As he makes clear in his article, Killeen (2005) turned his attention toward the derivation of a coefficient of replicability in response to what he perceived to be the existence of fundamental defects inherent to the inferential decision-making machinery employed by the psychological researcher. Reasonably enough, the chief object of his dissatisfaction was the practice of carrying out null hypothesis significance tests. Killeen (2005) noted that null hypothesis significance tests are "based on inferred sampling distributions, given a hypothetical value for a parameter" (2005, p. 345) and that decisions about parameters are made on the basis of p values. However, statisticians have developed a number of distinct logics for the testing of hypotheses about parameters, including the Fisherian, Neyman–Pearsonian, and Bayesian approaches. Whereas these logics are, in many respects, irreconcilable, the first two have been fused to form what Gigerenzer et al. (1989) call the *hybrid approach* (see also Chow, 1996), and it is this approach that is dominant in the testing of hypotheses within psychology. Because the Fisherian, Neyman–Pearsonian, Bayesian, and hybrid approaches have complicated linkages to issues that arise in a consideration of p_{rep} , it is worthwhile to briefly review each.

Michael Maraun, Department of Psychology, Simon Fraser University;
Stephanie Gabriel, Department of Statistics and Actuarial Science, Simon Fraser University.

We wish to thank J. Don Read for the suggestion to investigate p_{rep} .

Correspondence concerning this article should be addressed to Michael Maraun, Department of Psychology, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6. E-mail: maraun@sfu.ca

We follow Killeen (2005) in using a control group/experimental group design for purposes of exemplification. Thus, let E and C stand for experimental and control conditions, respectively, and let $X_j, j \in \{E, C\}$, represent the dependent variate conditional on each of E and C , and with conditional distributions $X_j \sim N(\mu_j, \sigma^2)$, $j \in \{E, C\}$. The aim is to decide whether the treatment has an impact on the dependent variate, a question standardly paraphrased as whether the two population means, μ_E and μ_C , are equal.¹ Because this question pertains to parameters, the problem is inferential in nature, and decision making will be based on samples of size n_E and n_C that are drawn from E and C .

Fisher was largely responsible for marrying experimental design and analysis to statistical analysis (Gigerenzer et al., 1989). Given the setting of a control group/experimental group design, the aim of a Fisherian statistical analysis is to test the null hypothesis H_0 : $\mu_E = \mu_C$. To this end, a statistic must be found whose sampling distribution is known under the condition that H_0 is true, and, in

this particular case, the classical choice is $t = \frac{(M_E - M_C)}{s_p \sqrt{\frac{1}{n_E} + \frac{1}{n_C}}}$, in

which M_E and M_C are the two sample means and s_p is the pooled within-sample standard deviation. If H_0 happens to be true, t has a central t distribution on $(n_E + n_C - 2)$ degrees of freedom. On the basis of the single samples taken from each of E and C , a single realization of t , called t_{obs} , is produced. The *significance level* (called now the p value or p_{obs}) of the test statistic is defined as $P(|t| > |t_{\text{obs}}| \mid H_0 \text{ true})$ and is used to quantify the discordance of the data with the null hypothesis. Although the p value was a legacy of the theory of outlier detection employed by astronomers, in particular, Benjamin Peirce's 1852 rule for the rejection of outlying astronomical observations (Gigerenzer et al., 1989), Fisher himself was uneasy with its employment and acknowledged that it was "not very defensible save as an approximation" (Fisher, 1956, p. 66). According to Fisher, whereas the p value quantifies evidence against H_0 (the smaller the value, the greater the evidence against H_0), it cannot be validly employed to support the conclusion that H_0 is true. Although Fisher entertained the idea that a single p value of sufficiently small magnitude might be taken as grounds for (provisionally) rejecting H_0 , he was actually a staunch believer in the necessity of carrying out replication attempts (Fisher, 1935). To him, the communication of results was of fundamental importance, and he believed that all p values should be published, thereby allowing the research community the opportunity to consider the cumulative evidence against H_0 . In Fisher's view, the p value expressed an evidential relationship between an individual sample and H_0 on the basis of which the researcher formed a mental attitude toward H_0 (Fisher, 1956).

Neyman and Pearson (1928/1967 and 1933, for example), who viewed certain features of the Fisherian approach as logically unfounded, replaced Fisher's single null hypothesis with a null and alternative hypothesis pair, the respective members of which were H_0 and H_1 . They viewed the aim of a hypothesis test as being that of making an optimal decision about which of the possible states of nature, H_0 or H_1 , is the case. On the Neyman-Pearsonian account, such decision making risks two types of error: the Type I error, rejecting H_0 when it is, in fact, true, and the Type II error, retaining H_0 when it is false. An optimal decision-making strategy is one in which $\alpha = P(\text{Type I error})$ is made small and $\beta(\delta, \alpha) = P(\text{retain } H_0 \mid \alpha, H_0$

false by $\delta)$ is made small for fixed α and values of δ the detection of which are of interest. For fixed α , the latter aim is realized by choosing the sample size to be such that $\text{Power}(\delta, \alpha) = 1 - \beta(\delta, \alpha)$ is made acceptably large for values of δ the detections of which are of interest. Through a consideration of error rates, and the costs inherent to their control, a formal decision rule is created, and, on the basis of single samples drawn from each of X_E and X_C , the researcher decides whether he or she will *act as if* H_0 is the case or H_1 is the case. When a decision about which of H_0 and H_1 is made on the basis of an individual test, the researcher is instantly either correct or incorrect and cannot know which. Confidence in decision making does not come from the results yielded by a particular set of data but from the long-run error control properties of the test procedure under which decisions are made.

Fisher's approach to inference was, in part, a response to the Bayesian approach. In the Bayesian approach, probabilities are used to express the degree of belief in, or plausibility of, hypotheses about parameters (see, e.g., Gelman, Carlin, Stern, & Rubin, 2004). The focal point of Bayesian inference regarding the hypotheses H_0 and H_1 is the posterior probability that expresses the degree of plausibility that the data confer upon the null hypothesis. From Bayes's theorem, the posterior probability of H_0 is as follows:

$$P(H_0 \text{ true} \mid |t| > |t_{\text{obs}}|) = \frac{P(|t| > |t_{\text{obs}}| \mid H_0 \text{ true})P(H_0 \text{ true})}{P(|t| > |t_{\text{obs}}|)}.$$

On the Bayesian account, the researcher should favor H_0 if $P(H_0 \text{ true} \mid |t| > |t_{\text{obs}}|) > 0.5$. However, in order to compute this posterior, the researcher must possess knowledge of the prior probability of the null being true. Frequentist statisticians such as Fisher, Neyman, and Pearson deny that these prior distributions can be given a coherent meaning within the probability calculus, whereas Bayesian statisticians reject the necessity of frequentist interpretations, taking priors to be nonfrequentist quantifications of the researcher's prior degree of belief in each hypothesis (Gelman et al., 2004; Silvey, 1970). Interestingly, although Fisher was a critic of nonfrequentist Bayesian posterior probabilities, he nonetheless invented a highly controversial Bayesian-like approach called *fiducial inference* (see Hannig, 2006), a topic that arises in Macdonald's (2005) commentary on Killeen's (2005) article.

The hybrid approach is a mixture of Fisherian and Neyman-Pearsonian logic (Chow, 1996). In statistical analyses of the hybrid sort, the researcher will, among other things: (a) follow Neyman and Pearson (1967) in attending to error rates; (b) replace Neyman and Pearson's frequentist interpretation of these rates as bearing on the quality of the test procedure with Fisherian styled claims such as "the null hypothesis was rejected, $p < .1$," which express a mental attitude toward H_0 ; and (c) adhere to Fisher's (1935) prohibition on the drawing of conclusions on the basis of nonsignificant results. Strictly speaking, the term *null hypothesis significance test* applies to hypothesis tests carried out within a Fisherian framework. However, a careful consideration of Killeen's (2005) criticisms of inference makes it clear that they apply equally well

¹ It should be noted that Fisher (1956) was careful to distinguish between a hypothesis stated in terms of parameters and the scientific hypothesis having to do with the causal efficacy of the treatment. The former is only a convenient stand-in for the latter.

to the Fisherian, Neyman–Pearsonian, and hybrid approaches. Thus, we take the term *null hypothesis significance test*, or *significance test* for short, to designate inferential hypothesis tests carried out within any of these three frameworks.

According to Killeen (2005), significance tests are problematic for two reasons. First, the decisions they render about hypotheses are based on the p value. Along with many others (e.g., Berger & Selke, 1987; Nickerson, 2000), Killeen considers the p value to be a suboptimal data-based measure of the plausibility of a given null hypothesis (an opinion held, as a matter of fact, by Fisher, 1956, himself). Killeen's belief is that what would be truly useful to the researcher are Bayesian posterior probabilities of the form $P(H_0 \text{ true} \mid \text{data})$. However, the calculation of a posterior probability requires that the researcher be in possession of a prior distribution, and priors "are largely unknowable" (Killeen, 2005, p. 345). Thus, Killeen (2005) concludes, "Significance tests without priors are the 'flaw in our method'" (p. 345). Second, Killeen (2005) views psychology as misdirected in following the lead of Fisher in placing such great importance on the acquisition of knowledge pertaining to parameters: "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" (2005, p. 345). He quotes, with approval, Geisser's (1992) lamentation that the cost of this preoccupation has been that the field of psychology has paid little attention to "inference about observables" (p. 1) and concludes that "our unfortunate historical commitment to significance tests forces us to rephrase these good questions [hypotheses of the existence of causal relations] in the negative, attempt to reject these nullities, and be left with nothing we can logically say about the questions" (2005, p. 346).

Killeen (2005) acknowledges that the defects he identifies have been discussed for some time and concludes that "when so many experts disagree on the solution, perhaps the problem itself is to blame" (p. 345). Thus, Killeen's (2005) most critical step is to attempt to provide "an alternative, one that shifts the argument by offering 'a solution to the question of replicability'" (p. 346). This alternative involves expressing the signal-to-noise information generated by an experiment as a coefficient of replicability. Killeen (2005) states that "Greenwald, Gonzalez, Guthrie, and Harris (1996) reviewed the NHST [null hypothesis significance test] 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' , is equal to the population effect size, δ " (p. 346). However, the focus of Greenwald et al. on the probability of replicating a rejection of a null hypothesis "replicates the dilemma of significance tests" (cited in Killeen, 2005, p. 346). Killeen's desire to move the discipline away from the notion of significance led him to attempt to derive quite a different probability, namely, the probability of obtaining "an effect of the same sign as that found in an original experiment" (2005, p. 346). The fruit of his attempt to derive this probability is his p_{rep} , a coefficient that he claims overcomes the defects inherent to significance tests by capturing "traditional publication criteria for signal-to-noise ratio, while avoiding parametric inference and the resulting Bayesian dilemma" (Killeen, 2005, p. 345).

Killeen's (2005) Derivation of p_{rep}

In subsequent sections of this article, we establish that p_{rep} is not equivalent to the probability that Killeen (2005) set out to

derive, that being the probability of obtaining an effect of the same sign as that found in an original experiment. Killeen claimed that, in p_{rep} , he had produced a version of this target probability that was not dependent on the unknown parameter δ . However, as we show below, the illusion that the dependency of the target probability on δ had been eliminated is traceable to an error present in Killeen's derivation of p_{rep} . It is, therefore, essential to begin with a step-by-step summary of Killeen's derivation of p_{rep} .² According to Killeen:

1. The statistic $d' = \frac{M_E - M_C}{s_p}$ is the *sample effect* and is the sample analogue of the *population effect* $\delta = \frac{\mu_E - \mu_C}{\sigma}$.
2. As $n_E + n_C$ becomes large, d' will come to be normally distributed with a mean of δ and a variance of $\sigma_d^2 \approx \frac{(n_E + n_C)^2}{n_E n_C (n_E + n_C - 4)}$. That is to say, $d' \xrightarrow{\text{Lim}} N(\delta, \sigma_d^2)$.³
3. Let there be an original experiment, results pertaining to which are subscripted with a 1, and a hypothetical replication attempt subscripted with a 2. A replication has occurred if the hypothetical replication attempt yields "an effect of the same sign as that found in the original experiment" (Killeen, 2005, p. 346). Thus, if $d'_1 > 0$, a replication has occurred just when $d'_2 > 0$.⁴
4. Now, "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 " (Killeen, 2005, p. 346). The calculation of this probability requires that the researcher know the value of δ . Unfortunately, "we do not know the value of the parameter δ and must therefore eliminate it" (Killeen, 2005, p. 346).
5. Define the sampling errors $\Delta_1 = d'_1 - \delta$ and $\Delta_2 = d'_2 - \delta$. It follows from Step 2, that $E(\Delta_1) = E(\Delta_2) = 0$ and $V(\Delta_1) = V(\Delta_2) = \sigma_d^2$.
6. It follows from Step 5 that $\delta = d'_1 - \Delta_1$, $\delta = d'_2 - \Delta_2$, $d'_1 = \delta + \Delta_1$, and $d'_2 = \delta + \Delta_2$.
7. Because a replication has occurred when $d'_2 > 0$, it follows from Step 6, $d'_2 = \delta + \Delta_2$, that a replication has occurred when $\delta + \Delta_2 > 0$. However, also from Step 6,

² The organization of our summary of Killeen's derivation in terms of lettered points is our own.

³ Henceforth, we symbolize asymptotic results such as $d' \xrightarrow{\text{Lim}} N(\delta, \sigma_d^2)$ using the more compact notation $d' \rightarrow N(\delta, \sigma_d^2)$.

⁴ Without loss of generality, we only consider the case in which $d'_1 > 0$ and, hence, in which a replication has occurred when $d'_2 > 0$. The logical and mathematical principles that apply to this case are identical for the case in which $d'_1 < 0$ (and, hence, in which a replication has occurred when $d'_2 < 0$).

$\delta = d'_1 - \Delta_1$. Thus, it can be concluded that a replication has occurred when $(d'_1 - \Delta_1) + \Delta_2 > 0$.

8. From representation $d'_2 = (d'_1 - \Delta_1) + \Delta_2$, the fact that “the expectation of each sampling error is 0 with variance σ_d^2 ” (Killeen, 2005, p. 347), and the fact that d'_1 and d'_2 are independent replications, it follows that “the variances add, so that $d'_2 \sim N(d'_1, \sigma_{d_r}^2)$, with $\sigma_{d_r}^2 = \sqrt{2}\sigma_d$ ” (Killeen, 2005, p. 347).
9. Thus, if $d'_1 > 0$, the probability of obtaining an effect of the same sign as that found in the original experiment is $P(d'_2 \geq 0)$, and it follows from the deduced distributional result $d'_2 \sim N(d'_1, \sigma_{d_r}^2)$ that this latter probability is equal to

$$p_{\text{rep}} = \int_{-\infty}^{\frac{d'_1}{\sqrt{2}\sigma_d}} N(0, 1), \tag{1}$$

which is the area under the standard normal distribution beneath the point $\frac{d'_1}{\sqrt{2}\sigma_d}$ (Killeen, 2005, p. 347), as is illustrated in Figure 1.

Problems with Killeen’s (2005) p_{rep} Coefficient

Having detailed Killeen’s (2005) derivation of p_{rep} , we will, in this section: (a) establish that p_{rep} is not the probability that Killeen believed that he had derived, that being the probability of obtaining an effect of the same sign as that found in an original experiment. Killeen was, in fact, incorrect in his assessment that the dependency of this target probability on the unknown parameter δ could be eliminated, and the illusion that he had succeeded in eliminating this dependency is attributable to mathematical error; (b) establish that the coefficient p_{rep} is simply a quasi-power coefficient; and (c) provide several examples that illustrate the difference between the quasi-power coefficient p_{rep} and the true in vacuo replication probability $P(d'_2 > 0 \mid d'_1 > 0)_{\text{iv}}$ (where iv = in vacuo) that Killeen set out to derive.

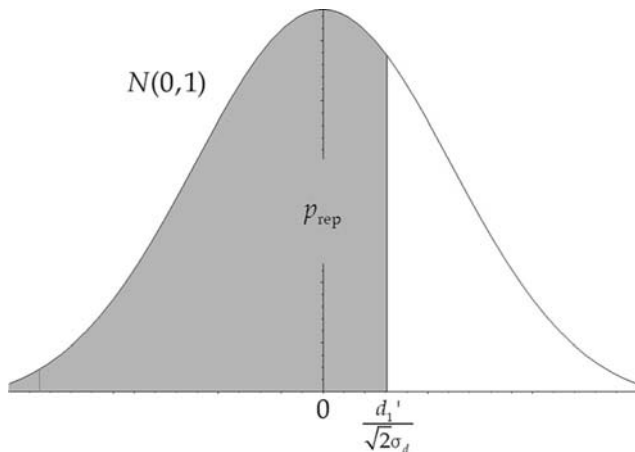


Figure 1. The area under the normal curve corresponding to p_{rep} .

The Coefficient p_{rep} Is Not the Probability Killeen (2005) Sought to Derive

Up to and including Step 7, there is nothing incorrect about Killeen’s (2005) derivation. Steps 1 through 7 spell out the assumptions on which Killeen’s derivation is based. These assumptions are the following:

1. d'_1 and d'_2 are each normally distributed;
2. d'_1 and d'_2 have the same mean, δ , and variance, σ_d^2 ;⁵
3. d'_1 and d'_2 are statistically independent.

The fact that Killeen took Assumptions 1 through 3 as the basis for his derivation indicates that by “replication,” he meant “replication in vacuo.” In an in vacuo replication scenario, replication attempts are imagined as being set within an unchanging world (in particular, within an idealized world in which whatever happens in the original experiment does not in any way influence what will happen in a replication attempt). Clearly, then, in an in vacuo replication scenario, the researcher does not actually carry out a replication attempt. Instead, a random variate d'_2 is invented and assigned distributional properties. This random variate stands for a theoretical infinity of sample effects imagined as having been produced in vacuo. Because d'_1 and d'_2 are defined on the same population of individuals and, hence, are jointly distributed, Assumptions 1 through 3 are equivalent to the single distributional claim that d'_1 and d'_2 have an asymptotic bivariate normal distribution with mean vector $\begin{pmatrix} \delta \\ \delta \end{pmatrix}$ and covariance matrix $\begin{pmatrix} \sigma_d^2 & 0 \\ 0 & \sigma_d^2 \end{pmatrix}$ or, equivalently, that $d'_1 = \delta + \Delta_1$ and $d'_2 = \delta + \Delta_2$, in which Δ_1 and Δ_2 have an asymptotic bivariate normal distribution, with mean vector $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} \sigma_d^2 & 0 \\ 0 & \sigma_d^2 \end{pmatrix}$. The zero off-diagonal elements of these covariance matrices indicate that the covariance between d'_1 and d'_2 is equal to zero; this condition is equivalent to Assumption 3 under the bivariate normality of d'_1 and d'_2 .

The error present in Killeen’s (2005) derivation occurs at Step 8 of the derivation. For the case in which $d'_1 > 0$, Killeen’s p_{rep} was supposed to be equal to $P(d'_2 \geq 0)$, in which d'_2 is a normally distributed random variate. As is well known, for the case of an arbitrary random variate w that is distributed as $N(\mu, \sigma^2)$, the probability that w will assume a positive value, $P(w \geq 0)$, is equal to the probability that a standard normal random variate z will assume a value less than or equal to $\frac{\mu}{\sigma}$. This latter probability,

$P\left(z \leq \frac{\mu}{\sigma}\right)$, is equal to $\int_{-\infty}^{\frac{\mu}{\sigma}} N(0, 1)$, the area under the standard normal distribution to the left of the point defined by the ratio of the mean and standard deviation of w . At Step 8 of the derivation, Killeen makes the deduction that $d'_2 \sim N(d'_1, 2\sigma_d^2)$,

⁵ In fact, Killeen makes an asymptotic argument in support of his claim that d'_1 and d'_2 are each normally distributed, and the homogeneity of the variances of d'_1 and d'_2 is a consequence of this asymptotic argument. Thus, technically speaking, neither Assumption 1 nor the equality of variance property is an assumption.

and this deduction asserts that the mean of d'_2 is equal to d'_1 and that the variance of d'_2 is equal to $2\sigma_d^2$. In p_{rep} , Killeen was attempting to produce $P(d'_2 \geq 0)$, and it follows from his deduction that $P(d'_2 \geq 0)$ should be equal to $\int_{-\infty}^{\frac{d'_1}{\sqrt{2}\sigma_d}} N(0, 1)$, the area under the standard normal distribution to the left of the point defined by the ratio of the mean and standard deviation that Killeen deduced for d'_2 .

However, Killeen's (2005) Step 8 deduction $d'_2 \sim N(d'_1, 2\sigma_d^2)$ that is the basis for p_{rep} is incorrect. It follows directly from Killeen's Assumptions 1 through 3 that the mean of d'_2 is not equal to d'_1 but rather to δ :

$$\begin{aligned} E(d'_2) &= E([d'_1 - \Delta_1] + \Delta_2) \\ &= E(d'_1) - E(\Delta_1) + E(\Delta_2) \\ &= \delta + 0 + 0 \\ &= \delta \end{aligned} \tag{2}$$

The variance of d'_2 is not equal to $2\sigma_d^2$ but rather to σ_d^2 :

$$\begin{aligned} V(d'_2) &= V([d'_1 - \Delta_1] + \Delta_2) \\ &= V(d'_1) + V(\Delta_1) + V(\Delta_2) - 2C(d'_1, \Delta_1) + 2C(d'_1, \Delta_2) \\ &\quad - 2C(\Delta_1, \Delta_2) \\ &= \sigma_d^2 + \sigma_d^2 + \sigma_d^2 - 2\sigma_d^2 + 2(0) + 2(0) \\ &= 3\sigma_d^2 - 2\sigma_d^2 \\ &= \sigma_d^2 \end{aligned} \tag{3}$$

Thus, given Killeen's (2005) stated assumptions, the distribution of d'_2 is actually $N(\delta, \sigma_d^2)$, and this distribution certainly *does* depend on unknown parameter δ . Thus, so too must the probability $P(d'_2 \geq 0)$. Coefficient p_{rep} does not depend on δ , but, contrary to what Killeen claimed, is certainly not a parameter-free version of the replication probability that he set out to derive. As we establish in the next section of the paper, p_{rep} is not a replication probability at all, but is, rather, a quasi-power coefficient. The illusion that Killeen had eliminated the dependency of $P(d'_2 \geq 0)$ on the unknown parameter δ was produced by his substitution of an incorrect mean (d'_1) in place of the true mean of d'_2 (i.e., δ) and his substitution of an incorrect variance ($2\sigma_d^2$) in place of the true variance of d'_2 (i.e., σ_d^2).

Killeen's incorrect distributional deduction $d'_2 \sim N(d'_1, 2\sigma_d^2)$ seems to have been a result of his changing, mid-derivation, the assumptions on which the derivation was based. As may be recalled, Killeen began the derivation under the stated Assumptions 1 through 3, which assert that d'_1 and d'_2 are statistically independent, normally distributed random variates with identical means (δ) and identical variances (σ_d^2). Assumptions 1 through 3 are equivalent to the single claim that $d'_1 = \delta + \Delta_1$ and $d'_2 = \delta + \Delta_2$, in which Δ_1 and Δ_2 have a bivariate normal distribution with mean vector $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} \sigma_d^2 & 0 \\ 0 & \sigma_d^2 \end{pmatrix}$. However, as the following simple derivations show, Killeen's deduction $d'_2 \sim N(d'_1, 2\sigma_d^2)$ would be correct only if d'_1 were a constant and Δ_1 and Δ_2 were random variates with mean vector $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} \sigma_d^2 & 0 \\ 0 & \sigma_d^2 \end{pmatrix}$:

$$\begin{aligned} E(d'_2) &= E([d'_1 - \Delta_1] + \Delta_2) \\ &= E(d'_1) - E(\Delta_1) + E(\Delta_2) \quad (\text{linearity property of} \\ &= d'_1 + 0 + 0 \quad \text{expectation operator} \\ &= d'_1 \quad \text{(mean of constant} \\ & \quad \text{is constant;} \\ & \quad E(\Delta_1) = E(\Delta_2) = 0 \\ & \quad \text{by definition)} \end{aligned} \tag{4}$$

$$\begin{aligned} V(d'_2) &= V([d'_1 - \Delta_1] + \Delta_2) \\ &= V(d'_1) + V(\Delta_1) + V(\Delta_2) \quad (\text{by independence of} \\ &= 0 + \sigma_d^2 + \sigma_d^2 \quad d'_1, \Delta_1, \text{ and } \Delta_2) \\ &= 2\sigma_d^2 \quad (V(d'_1) = 0; \\ & \quad V(\Delta_1) = V(\Delta_2) = \sigma_d^2) \end{aligned} \tag{5}$$

Clearly, the claim that d'_1 is a constant and Δ_1 and Δ_2 are random variates is in conflict with Killeen's (2005) stated Assumptions 1 through 3, and the result is an array of mathematical contradictions and, ultimately, a p_{rep} coefficient that is not equivalent to the replication probability that Killeen set out to derive. For example, if, as needed to make correct the deduction $d'_2 \sim N(d'_1, 2\sigma_d^2)$, d'_1 were a constant, then Killeen's distributional claim $d' \sim N(\delta, \sigma_d^2)$ would be incorrect: Constants cannot be normally distributed. Alternatively, if d'_1 were a constant, then it could not be true that $d'_1 = \delta + \Delta_1$, in which Δ_1 is a normally distributed random variate, with a mean of 0 and a variance of σ_d^2 . The latter distributional statement could not be true because, if both d'_1 and δ were constants, then, by the equation $d'_1 = \delta + \Delta_1$, so, too, would Δ_1 have to be. But in that case, Δ_1 could not be a normally distributed random variate with a mean of 0 and variance of σ_d^2 . Constants can neither be normally distributed nor have a nonzero variance.

In our opinion, the incorrect deduction that ruined Killeen's (2005) derivation resulted from his improper mathematical handling of the notion of d'_1 having been observed. Killeen seems to have been of the view that when a realization d'_1^* is taken on random variate d'_1 that is jointly distributed with random variate d'_2 , the act of producing this realization turns random variate d'_1 into a constant. (This would explain how he came to change, mid-derivation, the assumptions on which the derivation was based.) However, this view is certainly mistaken. The act of taking a realization of random variate d'_1 does not turn d'_1 into a constant.⁶ Once a realization d'_1^* of random variate d'_1 is taken, the information this realization contains about random variate d'_2 is encoded in the conditional distribution of d'_2 given that $d'_1 = d'_1^*$. The conditional distribution of d'_2 given that $d'_1 = d'_1^*$ is a weighted "slice" of the bivariate distribution of d'_1 and d'_2 , the slice determined by d'_1^* . If d'_2 is dependent on d'_1 , then the conditional distribution of d'_2 given that $d'_1 = d'_1^*$ will be more informative about d'_2 than is the unconditional distribution of d'_2 .

According to Killeen's (2005) stated assumptions, d'_1 and d'_2 are jointly distributed random variates, and this means that the probability that Killeen needed to derive was a conditional probability:

⁶ Note that a constant can be characterized as a degenerate random variate, and taking a realization of a random variate does not turn it into a degenerate random variate.

If, in the initial experiment, $d'_1 = d_1^* > 0$, then the required probability is $P(d'_2 > 0 | d'_1 > 0)$; if, in the initial experiment, $d'_1 = d_1^* < 0$, then the required probability is $P(d'_2 < 0 | d'_1 < 0)$. Because these conditional probabilities are defined under a very particular conception of replication, namely, the *in vacuo* scenario, we henceforth symbolize them as $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$ and $P(d'_2 < 0 | d'_1 < 0)_{\text{iv}}$.⁷ If two random variates X and Y are statistically independent, then the conditional distribution of Y given X reduces to the unconditional distribution of Y .⁸ Thus, under the statistical independence assumption (i.e., Assumption 3) of the *in vacuo* scenario, $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$ reduces to $P(d'_2 > 0)_{\text{iv}}$. Because the distribution of d_2 is $N(\delta, \sigma_d^2)$, $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}} = P(d'_2 > 0)_{\text{iv}}$ is then equal to

$$\int_{-\infty}^{\frac{\delta}{\sigma_d}} N(0, 1), \tag{6}$$

which is the area under the standard normal distribution beneath the point $\frac{\delta}{\sigma_d}$ (see also Doros & Geier, 2005). As Equation 6 makes clear, the *in vacuo* replication probability $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$, the derivation of which was Killeen's (2005) aim, most certainly does depend upon the unknown parameter δ .

The Statistic p_{rep} Is Simply a Quasi-Power Coefficient

We established in the previous section that, given Killeen's (2005) Assumptions 1 through 3 that define the *in vacuo* replication scenario, p_{rep} is not equivalent to the replication probability $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$ that he set out to derive. As it is already the case that researchers are urged to include p_{rep} rather than the p value in their submissions to *Psychological Science*, it is somewhat urgent that we come to an understanding of exactly what it is that researchers are calculating when they calculate p_{rep} . We now establish that p_{rep} is not a replication probability at all, but rather a quasi-power coefficient.

Imagine a researcher who, in an initial experiment, wishes to test the hypothesis pair $H_0: \delta_1 = 0$ and $H_1: \delta_1 \neq 0$. The researcher decides to run a size α test, and because $d'_1 \rightarrow N(\delta_1, \sigma_d^2)$, the critical value c on which the decision rule "reject H_0 if $|d'_1 \text{ obs}| > c$ " is based is that value for which $1 - \int_{-c(\alpha)}^{c(\alpha)} N(0, \sigma_d^2) = \alpha$. Now imagine that, on the basis of a realization d_1^* of d'_1 , the researcher makes the decision to reject $H_0: \delta_1 = 0$ and, because of this rejection, decides to attempt a replication of the experiment. That is, he or she decides to draw a second sample and test the hypothesis pair $H_0: \delta_2 = 0$ and $H_1: \delta_2 \neq 0$ under the distributional framework $d'_2 \rightarrow N(\delta_2, \sigma_d^2)$.

As part of the research design supporting the planned replication attempt, a power analysis is carried out. The power function

$$\text{Power}(\delta^*, \alpha) = 1 - \int_{-c(\alpha)}^{c(\alpha)} N(\delta^*, \sigma_d^2)$$

is a property of the test procedure. For any possible departure δ^* , $\text{Power}(\delta^*, \alpha)$ quantifies the likelihood that the test procedure will detect a departure of δ^* should it actually exist when the size α procedure is employed. When one assesses $\text{Power}(\delta^*, \alpha)$ for a

particular possible departure δ^* , one is not making an assumption that δ^* is the value that parameter δ_2 will assume at the time of testing. At the time of testing, δ_2 will assume whatever value it will assume, and the entire aim in conducting the test procedure is to make a solid inference about this unknown value. In undertaking a power analysis, the researcher assesses how the test procedure will do *if* it encounters δ^* when it is employed and perhaps even makes alterations to the test procedure if he or she deems this assessed performance to be inadequate. However, at the moment that the test procedure is employed, nature will be as it is, the test procedure either will or will not yield the correct decision about δ_2 , and there is nothing that the test's power function can say about it.

No test procedure can deliver unimpeachable sensitivity of detection over the full range of values that can be assumed by δ_2 . The researcher must nominate a few values of δ_2 whose detection is of paramount importance. In the absence of any other evidence as to what the value of δ_2 will be, it would be perfectly reasonable for the researcher to calculate Power (d_1^*, α). After all, d_1^* was calculated on the sample drawn in the first experiment, and this sample is thus far the only one that has been drawn. There is, of course, no necessity that δ_2 will assume a value that is close to d_1^* when the replication attempt is actually carried out. The value d_1^* produced in the initial experiment might have been a poor estimate of δ_1 , or it may be the case that reality will change before the replication attempt is carried out, the result being that δ_1 and δ_2 are markedly unequal. Nonetheless, in the absence of any additional evidence, d_1^* is the best guess as to the value that δ_2 will assume when the replication attempt is carried out, and Power (d_1^*, α) is the probability that the test procedure will detect this departure *should* it occur in the replication attempt. This particular hypothetical, produced by inserting the observed sample effect into the power function, is sometimes called *observed power* (Onwuegbuzie & Leech, 2004; O'Keefe, 2007).

If p_{rep} were not based on the incorrect variance $2\sigma_d^2$ but rather on the correct variance σ_d^2 , it would be equal to

$$\int_{-\infty}^{d_1^*} N(0, 1) \tag{7}$$

and, because $d'_2 \rightarrow N(\delta_2, \sigma_d^2)$, Expression 7 can be read off directly as the probability that d'_2 will be greater than 0 given that its mean, the parameter δ_2 , happens to be equal to d_1^* . However, p_{rep} is based on an incorrect variance, and therefore, rather than being equal to Expression 7, it is a lower bound to Expression 7:

$$p_{\text{rep}} = \int_{-\infty}^{\frac{d_1^*}{\sqrt{2}\sigma_d}} N(0, 1) \leq \int_{-\infty}^{\frac{d_1^*}{\sigma_d}} N(0, 1). \tag{8}$$

It may be concluded, then, that p_{rep} is a lower bound to the probability that d'_2 will be greater than 0, given that the unknown parameter δ_2 happens to be equal to d_1^* . Equivalently, it is a lower bound to the proportion of an imagined infinity of samples that

⁷ Because $P(d'_2 > 0 | d_1^* > 0)_{\text{iv}}$ and $P(d'_2 < 0 | d_1^* < 0)_{\text{iv}}$ are logically of a piece, to simplify our presentation we discuss only the former.

⁸ Similarly, the conditional distribution of X given Y reduces to the unconditional distribution of X .

would yield a positive sample effect given that δ_2 happens to be equal to d_1^* .

Coefficient p_{rep} is not a replication probability. Killeen's (2005) substitution of the realization d_1^* in place of the unknown parameter δ_2 makes p_{rep} a hypothetical akin to the power coefficients that researchers compute as a matter of course. Each of the values returned by Power (δ^*, α) is a hypothetical, equal to the probability of rejecting H_0 if nature happens to be a certain way when the size α test procedure is employed. Allowing δ^* to run over its range produces an infinite-sized class of hypotheticals (the points that form a power curve). Likewise, allowing d' , in the function

$$\int_{-\infty}^{d'} \frac{d'}{\sqrt{2\sigma_d}} N(0, 1),$$

to run over its range produces an infinite-sized class of hypotheticals. If one inserts d_1^* into Power (δ^*, α), one produces observed power, the probability that the size α test procedure will yield a rejection of H_0 given that δ_2 happens to be equal to d_1^* when the procedure is employed. If one inserts d_1^* in place of d' in the function

$$\int_{-\infty}^{d_1^*} \frac{d'}{\sqrt{2\sigma_d}} N(0, 1),$$

one produces p_{rep} , a lower bound to the probability that the procedure will yield a positive sample effect given that δ_2 happens to be equal to $d_1^* > 0$ when the procedure is employed. In fact, it can easily be proven that p_{rep} is a monotone-increasing function of observed power: Increases in observed power correspond to increases in the value of p_{rep} . Thus, the coefficient p_{rep} can reasonably be called a *quasi-power coefficient*.

In contrast to hypotheticals such as Power (δ^*, α), observed power, and p_{rep} , there is only one in vacuo probability of obtaining an effect of the same sign as that found in an original experiment. This probability, symbolized as $P(d_2' > 0 | d_1' > 0)_{\text{iv}}$, is equal to the proportion of samples produced in vacuo that contain a positive sample effect (i.e., for which $d_2' > 0$). The value it assumes in a given context is determined not by known properties of the sample but rather by the value assumed by the parameter δ . $P(d_2' > 0 | d_1' > 0)_{\text{iv}}$ is not a hypothetical but rather describes a state of nature that is unknown to the researcher. Figures 2, 3, and 4 illustrate the differences between $P(d_2' > 0 | d_1' > 0)_{\text{iv}}$, the replication probability Killeen (2005) set out to derive, the quasi-power coefficient p_{rep} he inadvertently produced, and observed power, the hypothetical that is p_{rep} 's close relative.

Examples

For the two populations E and C , let $\mu_E = 20$, $\mu_C = 18$, and $\sigma = 4.35$, so that $\delta = .46$. Under random sampling of $n_E = 10$ and $n_C = 10$ from E and C , respectively, it is approximately the case that $d_1' \sim N(\delta, \sigma_d^2)$ (Killeen, 2005), so that, under the current scenario, d_1' is distributed approximately as $N(.46, .25)$.⁹ Imagine, now, a researcher drawing samples of size 10 from each of E and C and calculating a d_1^* of 1.58, unaware, of course, of the fact that $d_1' \sim N(.46, .25)$. Using as input the realization 1.58, the researcher calcu-

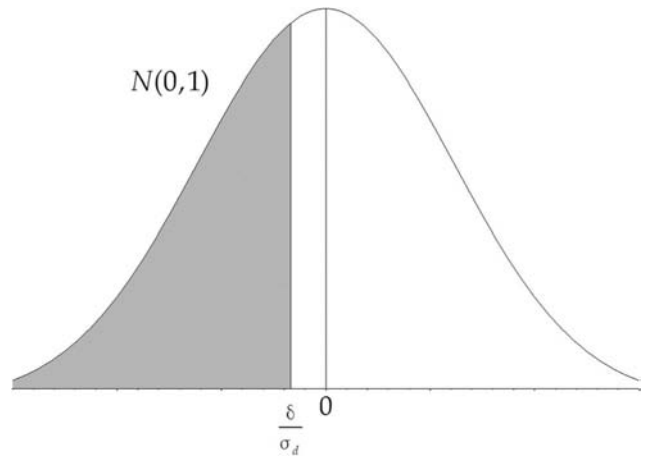


Figure 2. The true in vacuo replication probability $P(d_2' > 0 | d_1' > 0)_{\text{iv}}$.

lates a p_{rep} of .987. What does this value indicate? It certainly is not equal to the probability of obtaining, in a replication attempt, an effect of the same sign as that found in this original experiment, for p_{rep} is simply a quasi-power coefficient. A value of .987 for p_{rep} means that if an infinity of samples were produced under a population effect that *happened to be equal to 1.58* (the sample effect observed in the experiment), then at least 98.7% of these samples would yield a positive sample effect. The researcher could just as easily have calculated any other hypothetical, such as the proportion of positive sample effects that would be yielded by an infinity of samples under the condition that δ_2 *happened to be equal to $-.78$* (the answer being .13). Alternatively, if the researcher desired a hypothetical that contained exactly the same information as p_{rep} , he or she could calculate observed power, the probability that the test procedure will yield a rejection of H_0 given that δ_2 happens to be equal to d_1^* when the procedure is employed. If α were to be set to .01, observed power would be equal to .61.

Yet, the probability that Killeen sought to derive, $P(d_2' > 0 | d_1' > 0)_{\text{iv}}$, is not a hypothetical but rather a function of the unknown parameter δ . Under the current scenario, in which δ is equal to .46, $P(d_2' > 0 | d_1' > 0)_{\text{iv}}$ is equal to .82, meaning that if an infinity of samples were produced in vacuo, 82% of these samples would contain a positive sample effect. In this case, the true probability of an in vacuo replication is considerably lower than the value of p_{rep} , but it is easily shown that there does not exist a systematic relationship between the values assumed by these quantities in a particular context. Sometimes $P(d_2' > 0 | d_1' > 0)_{\text{iv}}$ will be lower than p_{rep} and sometimes it will be higher than p_{rep} . There does exist a systematic relationship between the hypotheticals p_{rep} and observed power, that being that they are monotone increasing functions of each other.

If, under an identical sampling scheme, δ happened instead to be equal to 1.15, and the sample drawn yielded a d_1^* of .09, p_{rep} would then be equal to .55, observed power would be equal to .005, and the true in vacuo replication probability, the propor-

⁹ Recall Result 2 from the section, Killeen's (2005) Derivation of p_{rep} section.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

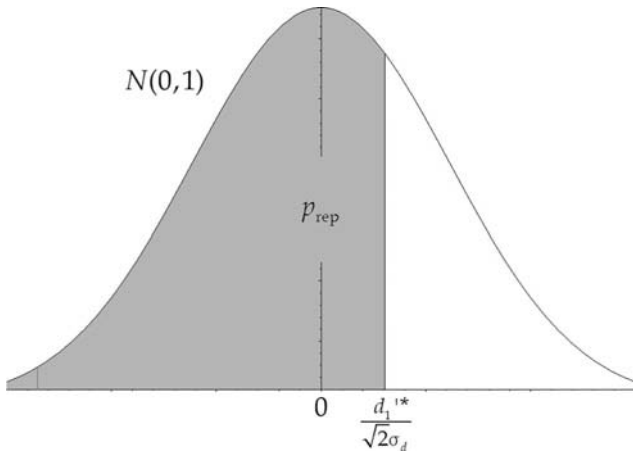


Figure 3. The quasi-power coefficient, p_{rep} .

tion of samples produced in vacuo that contain positive sample effects, would be equal to .99. Finally, if δ happened to be equal to 1.15 and the sample drawn yielded a d_1^{*} of -0.2 , p_{rep} would then be equal to .61, observed power would be equal to .007, and the true in vacuo replication probability $P(d_2' < 0 | d_1' < 0)_{\text{iv}}$, which in this case is equal to the probability of drawing a second sample effect whose sign does not agree with the population effect (some might call this the probability of a second anomalous result), would be equal to .01. Killeen (2005) believed that he had succeeded in eliminating the dependency of $P(d_2' < 0 | d_1' < 0)_{\text{iv}}$ on δ and that this probability was determined by the sample realization d_1^{*} , in this case, -0.2 . If this had been true, then it would make sense to take as truth the sample realization -0.2 and to consider the proportion of in vacuo-produced samples whose effects agreed in sign with this realization. However, Killeen was mistaken: $P(d_2' < 0 | d_1' < 0)_{\text{iv}}$ is controlled by the population effect, 1.15 in this case, and it is this value that represents truth when the issue is the impact of the treatment. In the current scenario, the researcher was *unfortunate* to draw a sample that yielded the value -0.2 , for this value creates a mistaken impression about the treatment's true impact, which is moderate and positive. The last thing that the researcher would desire is a replication of this misleading sample event.

Discussion

Contrary to what Killeen (2005) claimed, p_{rep} is not a replication probability. Killeen's mathematics contained an error that gave the (faulty) impression that the dependency of the in vacuo replication probability $P(d_2' > 0 | d_1' > 0)_{\text{iv}}$ on the unknown parameter δ could be eliminated. In replacing the unknown parameter δ with the observed sample effect d_1^{*} in the function

$$\int_{-\infty}^{\frac{\delta}{\sqrt{2}\sigma_d}} N(0, 1),$$

Killeen produced p_{rep} , a hypothetical that is a lower bound to the probability that d_2' will be positive, given that the unknown pop-

ulation effect *happens to be equal to* $d_1^{*} > 0$. Further, there does not exist a systematic relationship between the value yielded by $P(d_2' > 0 | d_1' > 0)_{\text{iv}}$ and the value yielded by p_{rep} : Sometimes the former will be the larger of the two, and sometimes the latter will be the larger. Moreover, just as for the values of the power function $\text{Power}(\delta^*, \alpha)$, the researcher does not need to conduct an experiment in order to know the values assumed by the function

$$\int_{-\infty}^{\frac{d'}{\sqrt{2}\sigma_d}} N(0, 1)$$

as d' runs over its range. To offer p_{rep} as a description of the signal-to-noise properties of an experimental outcome is akin to offering observed power for the same purpose, and it would be perverse to offer observed power as the final statement about an experiment designed to render a decision about the efficacy of a manipulation. We doubt that researchers would have accepted p_{rep} as a replacement for the p value had they been told the truth about it.

Despite the lack of usability of p_{rep} , Killeen (2005), nevertheless, has performed a service in pushing the discipline to step beyond standard p -value-based inference and consider the general issue of replication probabilities. We believe, however, that certain of Killeen's remarks on the topic of replication probabilities are misleading and require correction so that the discipline may engage in a fruitful discussion of this issue. First, we propose that Killeen was incorrect in his suggestions that replication probabilities can be made parameter free (that, essentially, data can control the parametric sampling distribution of a statistic) and that a change in focus from classical hypothesis testing to replication probabilities amounts to a change in focus from parameters to observables.

Probabilities are simply long-run relative frequencies or, equivalently, proportions defined on an infinitely long sequence of outcomes of a system. But the notion of an infinitely long sequence is an abstraction that can only be characterized in terms of generating functions and their parameters. Replication probabilities cannot be freed from their dependency on parameters. This is why, in

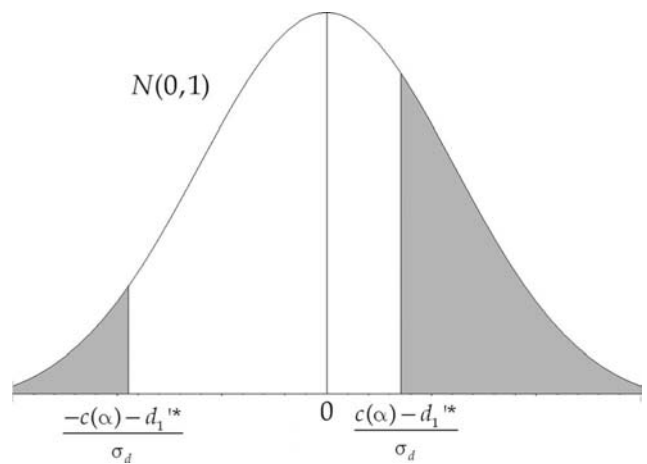


Figure 4. The area under the normal curve corresponding to observed power.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

their comment on Killeen's original article, Doros and Geier's (2005) attempts to derive p_{rep} on the basis of different sets of assumptions resulted in legitimate replication probabilities, each and every one of which featured parameters. It is also manifest in Killeen's (2005) derivation of p_{rep} , which features asymptotics (concerning the convergence of sample quantities to population quantities) and in vacuo replication scenarios (defined, in part, on the basis of the invariance of certain parameters). An interest in replication probabilities is a de facto interest in parameters.

It is odd that Killeen (2005) saw replication probabilities, such as $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$, as a potential improvement on p_{obs} , for a replication probability such as $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$ is not at all the same kind of quantity as p_{obs} . The value assumed by $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$ is determined by the parameter δ and, in order for the researcher to gain some insight into the value assumed by the parameter δ , he or she would have to engage in inferential decision making about that particular parameter. The coefficient p_{obs} , on the other hand, while admittedly suboptimal for the purpose, provides the researcher with a means of making inferential decisions about parameters. In replication probabilities such as $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$, Killeen has not offered the discipline a competitor to p_{obs} but has instead replaced one inferential problem, the classical need to make decisions about the value assumed by δ , with an equivalent one, the need to make inferences about the values assumed by replication probabilities.

Consider a scenario in which $n_E = 10$, $n_C = 10$, $\mu_E = 20$, $\mu_C = 18$, and $\sigma = .69$, so that $\delta = 2.9$. Because it is approximately the case that $d'_1 \sim N(\delta, \sigma_d^2)$, it follows that, in this scenario, d'_1 is distributed approximately as $N(2.9, 0.25)$. That δ is equal to 2.9 will, of course, be unknown to the researcher. Now, imagine that, on the basis of a single sample, the researcher obtains a d'_1 equal to $-.18$. Killeen (2005) would act as if parameters have been eliminated and that the (observable) sample realization $-.18$ has priority. It does not. The instant that the researcher invokes a population characterization, such as $d'_1 \sim N(\delta, \sigma_d^2)$, he or she is drawing a formal distinction between population and sample and implicitly asserting the priority of parameters. For it is the value assumed by the parameter δ , 2.9 in the current scenario, that controls the sampling distributions of statistics such as d'_1 and d'_2 and the value assumed by the replication probability $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$. So long as a formal distinction between population and sample is maintained and modeled, the natural definition of *treatment effect* is the parameter value 2.9.

Just as it is a misrepresentation to portray replication probabilities, such as $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$, as being parameter-free and to portray a focus on replication probabilities as a shift away from classical inference, so too is it confused to invoke the notion of "inference about observables," and to suggest, as does Killeen (2005), that a focus on replication probabilities amounts to a focus on observables. If by observable, one means a quantity calculated on the scores produced in an experiment, then knowledge of the scores allows the researcher to know the value of any observable of interest. The tools of inference are not needed when the analytic foci are observables. If Killeen is correct that psychology should supplant its interest in parameters with an interest in observables, then the tools of classical statistical inference can be set aside by the psychologist in favor of the tools of exploratory data analysis (Tukey, 1977). Certainly, however, the task of coming to know the value assumed by a replication probability such as $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$ is as inferential a task as there can be. If the discipline does, in fact, come to quantify the outcomes of exper-

iments through (noninferential) data-analytic explorations of observables, the need for replication probabilities, along with quantities such as p_{obs} and power, will have gone the way of the dinosaur.

In the second place, we suggest that Killeen (2005) misrepresented the kinds of claims that would be licensed by the researcher's knowledge of the value assumed by the in vacuo replication probability $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$. Consider a researcher who, in an experiment conducted in January, produces a d'_1 that is equal to 1.58, a value that inspires him or her to plan a replication attempt for July. That is to say, his or her plan is to conduct an experiment in July, the defining conditions of which are as similar as possible to those of the original experiment. Killeen's remarks make it sound as if the researcher's awareness of the value of $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$ would license statements about the likelihood that the July replication attempt will yield a positive sample effect. It would not. The in vacuo probability $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$ is defined under Assumptions 1 through 3 that define the in vacuo replication scenario. Whatever its virtues may be, this scenario does not provide a good description of how things work in nature.¹⁰ First, Assumption 2 of the in vacuo scenario asserts that the means of d'_1 and d'_2 , that is, δ_1 and δ_2 , are equal. However, if, as in the abovementioned scenario, d'_1 and d'_2 are sample effects defined under an initial experiment and a temporally removed replication attempt, there is absolutely no reason to believe that this condition will hold. In the real-world experimental contexts that psychologists face, change is the norm, this being the reason that latent growth curve models and time series models have become so very popular. Second, the very act of deciding to attempt a replication on the basis of a notable realization of d'_1 implies a violation of Assumption 3, which asserts the statistical independence of d'_1 and d'_2 .

In a real-world replication scenario involving an initial experiment and a temporally removed replication attempt, the relevant replication probability is the following:

$$P(d'_2 > 0 | d'_1 > 0) = \frac{P(d'_1 > 0 \cap d'_2 > 0)}{P(d'_1 > 0)},$$

in which $P(d'_1 > 0 \cap d'_2 > 0)$ is the empirical joint distribution function of d'_1 and d'_2 (i.e., the probability that d'_1 and d'_2 will both be greater than zero). The term *empirical joint distribution* is employed to distinguish $P(d'_1 > 0 \cap d'_2 > 0)$ from artificial, idealized distributional contexts such as the in vacuo scenario that is the basis for $P(d'_2 > 0 | d'_1 > 0)_{\text{iv}}$. $P(d'_1 > 0 \cap d'_2 > 0)$ is not in any way an idealization: Both its form and the values assumed by its parameters are whatever they are by virtue of the interaction of the features of the experiment that are

¹⁰ Nor does the idealization-heavy brand of inference that is classical hypothesis testing (famous for its invocation of independence of observations, population normality, etc.) provide accurate real-world descriptions. Once again, we urge the discipline to take seriously Killeen's (2005) preference for the study of observables by reducing its dependency on inferential questions and procedures and increasing its employment, as a basis for quantitative decision making, of noninferential data-analytic techniques (e.g., de Leeuw, 1988; Tukey, 1977), backed by jackknife- and bootstrap-based (see, e.g., Efron, 1979) stability analyses. Furthermore, Killeen was correct about the importance of replicability as a basis for quantifying the importance of a finding, and the data analytically based decision making we propose should be embedded within an ongoing program of attempted replications of findings.

under the researcher's control and those that are simply properties of nature itself. It is not knowledge of the value assumed by the in vacuo probability $P(d'_2 > 0 \mid d'_1 > 0)_{\text{iv}}$, but rather the real-world replication probability $P(d'_2 > 0 \mid d'_1 > 0)$, that would license the kinds of replicability statements that Killeen would like to make. To know the value assumed by $P(d'_2 > 0 \mid d'_1 > 0)$, however, is to know both the form of $P(d'_1 > 0 \cap d'_2 > 0)$ and the values assumed by its parameters. Unfortunately, the researcher will not know either of these required pieces of information. Perhaps the one thing that can be said about $P(d'_1 > 0 \cap d'_2 > 0)$ is that because the outcome of an initial experiment will influence the researcher's decision regarding whether to attempt a replication, it is likely that its form will be complicated.

Our view is that Killeen (2005) was correct that the issue of replicability should have a central role in researchers' assessments of the empirical results that they produce but that this role is properly ensured not through the employment of coefficients of replication but rather through more consistent efforts—chiefly in the form of a sustained program of replication attempts—to possess the capacity to replicate experimental outcomes. To possess such a capacity is to be able to specify the antecedent conditions necessary to bring about the outcome. Thus, a researcher's possession of this capacity is a powerful demonstration that he or she has identified the causal antecedents of a given outcome. A high value of a replication probability such as $P(d'_2 > 0 \mid d'_1 > 0)_{\text{iv}}$ is a characteristic of a sampling distribution and, hence, is in no way equivalent to the possession of this capacity.

References

- Berger, J. O., & Selke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of American Statistical Association*, 82, 112–122.
- Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. Thousand Oaks, CA: Sage.
- Cutting, J. E. (2005). Acknowledgement. *Psychological Science*, 20, 11–14. Retrieved May 2007 from <http://www.wiley.com/bw/submit.asp?ref=0956-7976&site=1>
- de Leeuw (1988). Multivariate analysis with linearizable regressions. *Psychometrika*, 53, 437–454.
- Doros, G., & Geier, A. B. (2005). Probability of replication revisited. *Psychological Science*, 16, 1005–1006.
- Efron, B. (1979). *Bootstrap methods: Another look at the jackknife*. *The Annals of Statistics*, 7, 1–26.
- Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society Series A*, 98, 39–54.
- Fisher, R. A. (1956). *Statistical methods in scientific inference*. Edinburgh, Scotland: Oliver & Boyd.
- 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.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall.
- Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge, England: Cambridge University Press.
- 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.
- Hannig, J. (2006). *On fiducial inference—The good, the bad and the ugly* [Tech. Report]. Colorado State University, Department of Statistics.
- Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16, 345–352.
- 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.
- Maccdonald, R. R. (2005). Why replication probabilities depend on prior probability distributions. *Psychological Science*, 16, 1007–1008.
- Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337.
- Neyman, J., & Pearson, E. (1967). On the use and interpretation of certain test criteria for purposes of statistical inference, Part I. In J. Neyman & E. Pearson, *Joint statistical papers*. Cambridge, England: Cambridge University Press. (Original work published 1928)
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241–301.
- O'Keefe, D. J. (2007). Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures*, 1, 291–299.
- Onwuegbuzie, A. J., & Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics*, 3, 201–230.
- Silvey, S. D. (1970). *Statistical Inference*. London, England: Chapman & Hall.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Received August 17, 2007

Revision received May 12, 2009

Accepted May 28, 2009 ■