Psychonomic Bulletin & Review
2010, 17 (2), 263-269
doi:10.3758/PBR.17.2.263

Notes and Comment

Replication is not coincidence: Reply to Iverson, Lee, and Wagenmakers (2009)

Bruno Lecoutre
CNRS and Université de Rouen, Rouen, France

and

Peter R. Killeen
Arizona State University, Tempe, Arizona

Iverson, Lee, and Wagenmakers (2009) claimed that Killeen's (2005) statistic p_{rep} overestimates the "true probability of replication." We show that Iverson et al. confused the probability of replication of an observed direction of effect with a probability of coincidence*—the probability that two future experiments will return the same sign. The theoretical analysis is punctuated with a simulation of the predictions of p_{rep} for a realistic random effects world of representative parameters, when those are unknown a priori. We emphasize throughout that p_{rep} is intended to evaluate the probability of a replication outcome after observations, not to estimate a parameter. Hence, the usual conventional criteria (unbiasedness, minimum variance estimator) for judging estimators are not appropriate for probabilities such as p and p_{rep}.*

Iverson, Lee, and Wagenmakers (2009; hereafter, ILW) claimed that Killeen's (2005) p_{rep} "misestimates the true probability of replication" (p. 424). But it was never designed to estimate what they call the true probability of replication (the broken lines named "Truth" in their Figure 1). We clarify that by showing that their "true probability" for a fixed parameter δ—their scenario—is the probability that the effects of *two future* experiments will agree in sign, given knowledge of the parameter δ. We call this the *probability of coincidence* and show that its goals are different from those of p_{rep}, the predictive probability that a future experiment will return the same sign as one already observed. ILW's "truth" has nothing to do with the "true probability of replication" in its most useful instantiation, the one proposed by Killeen (2005).

The "True Probability of Replication"

Statistical analysis of experimental results inevitably involves unknown parameters. Suppose that you have observed a positive standardized difference of $d_{\text{obs}} = 0.30$ between experimental and control group means having $n = 10$ subjects each.[1] You assume the usual normal model with an unknown true effect size δ and (for simplification) a known variance. What is the probability of getting again a positive effect in a replication ($d_{\text{rep}} > 0$)? If you are ready to assume a particular value for δ, the answer is trivial: It follows from the sampling distribution of d_{rep}, given this δ. The true probability of replication is the (sampling) probability $\varphi_{+|\delta}$ (a function of δ and n) that a normal variable with a mean of δ and a variance of $2/n$ exceeds 0: $\varphi_{+|\delta} = \Phi(\delta\sqrt{n/2})$. If you hypothesize that δ is 0, then $\varphi_{+|0} = 0.5$. Some other values, for different hypothesized δs, are $\varphi_{+|0.50} = 0.868$, $\varphi_{+|1.00} = 0.987$, $\varphi_{+|2.00} \approx 1$. These values do not depend on d_{obs}: It would not matter that $d_{\text{obs}} = 0.30$ or $d_{\text{obs}} = 1.30$. Of course, for reasons of symmetry, $\varphi_{+|-\delta} = 1-\varphi_{+|\delta}$.

What was novel about Killeen's (2005) statistic p_{rep} was his attempt to move away from the assumption of knowledge of parameter values, and the "true replication probabilities" $\varphi_{+|\delta}$ that can be calculated if you know them. The Bayesian derivation of p_{rep} involves no knowledge about δ other than the effect size measured in the first experiment, d_{obs}. This is made explicit by assuming an uninformative (uniform) prior before observations—hence, the associated posterior distribution for δ: a normal distribution centered on d_{obs} with a variance of $2/n$. To illustrate the nature and purpose of p_{rep}, consider the steps one must follow to simulate its value, starting with a known first observation:

Repeat the two following steps many times:

(1) generate a value δ from a normal(d_{obs},$2/n$) distribution;

(2) given this δ value, generate a value d_{rep} from a normal(δ,$2/n$);

and then compute the proportion of d_{rep} having the same sign as d_{obs}. Each *particular* value of d_{rep} is the realization of a particular experiment assuming a true effect size δ, and corresponds to a "true probability of replication" $\varphi_{+|\delta}$ (if $d_{\text{obs}} > 0$) or $1-\varphi_{+|\delta}$ (if $d_{\text{obs}} < 0$). But δ varies according to Step 1, which expresses our uncertainty about the true effect size, given d_{obs}. Hence, p_{rep} is a weighted mean of all the true probabilities of replication $\varphi_{+|\delta}$. This is the classic Bayesian posterior predictive probability (see, e.g., Gelman, Carlin, Stern, & Rubin, 2004). Explicit formulae for p_{rep} are given by Killeen (2005), and other references cited by ILW (2009). It is like a p value and a Bayesian posterior probability concerning a parameter, in that it is not designed to estimate a parameter but, rather, to be used as a decision aid (e.g., Killeen, 2006). Nonetheless, when the uncertainty about δ vanishes—when n is very large—p_{rep} tends to the true probability of replication $\varphi_{+|\delta}$. This is perfectly coherent.

P. R. Killeen, killeen@asu.edu

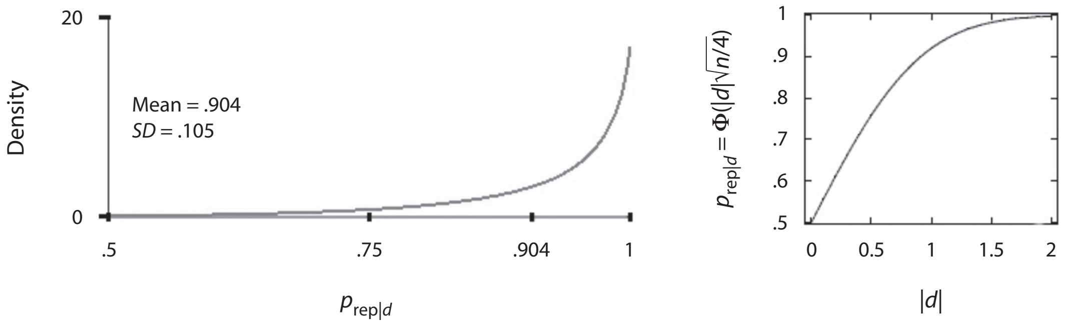

Figure 1. Sampling distribution of $p_{\mathrm{rep}|d}$ when $n = 10$ and the underlying true effect size is $\delta = 1$ (left panel) and the correspondence between $|d|$ and $p_{\mathrm{rep}|d}$ (right panel).

ILW (2009) asserted that the statistic p_{rep} is "a poor estimator," "biased and highly variable" of the "true probability of replication." Their analysis assumes "a fixed effect size (i.e., a δ value)" (p. 424) and a large number of imaginary repetitions of the experiment that are simulated under that hypothetical circumstance. It is a standard frequentist analysis that is intended to be done *before observations*, in order to study the sampling properties of a statistic. It must be contrasted with the Bayesian derivation of p_{rep} that is done *after* observations and assumes a fixed value d_{obs}: This leads to a unique value p_{rep}, Killeen's (2005), which, to avoid any ambiguity, should perhaps be denoted $p_{\mathrm{rep}|d\mathrm{obs}}$. Values of δ were not sampled as in the above Step 1. By contrast with the Bayesian approach, the frequentist approach considers all possible values of the sample standardized difference between means and, consequently, all possible values of p_{rep}. Both of these two quantities are treated as random variables. This requires different notations; here, we use d and $p_{\mathrm{rep}|d}$ to keep these separate.

For each simulated experiment, ILW (2009) computed the standardized difference between means, d, and its associated $p_{\mathrm{rep}|d} = \Phi(|d|\sqrt{n/4})$. This procedure generates the sampling distributions of these two statistics. Such a simulation can be fruitfully employed to illustrate the fact that in the normal case with known variance, d is a "good" (unbiased, minimum variance) estimator of δ: If you compute the moments of the sampling distribution of d generated from a very large number of repetitions, you will find mean$(d) = \delta$ and var$(d) = 2/n$ (with an approximation depending on the number of repetitions).

ILW (2009) applied the same approach to the statistic $p_{\mathrm{rep}|d}$ and computed the mean and standard deviation of its sampling distribution for fixed effect size values δ varying from 0 to 2. The results of their simulations for $n = 10$ in experimental and in control groups are shown in ILW's Figure 1. However, since there is a one-to-one correspondence (illustrated in the right panel of our Figure 1) between $p_{\mathrm{rep}|d}$ and $|d|$—$p_{\mathrm{rep}|d} = \Phi(|d|\sqrt{n/4})$—the exact sampling distribution of $p_{\mathrm{rep}|d}$ can be derived from the distribution of d by a simple change of variable. For instance, when the underlying true effect size is $\delta = 1$, we get the sampling distribution of $p_{\mathrm{rep}|d}$ (given $\delta = 1$) shown in the left panel of Figure 1. Instances of $p_{\mathrm{rep}|d}$ values vary from 0.5 (for $d = 0$) to 1 (for $|d| = 1$). We find mean$(p_{\mathrm{rep}|d}|\delta = 1) = 0.904$ and std$(p_{\mathrm{rep}|d}|\delta = 1) = 0.105$.[2] The mean of 0.904 is what is called "the mean performance of p_{rep}" by ILW in the caption of their Figure 1: It corresponds to their circular marker associated with the (true) effect size of 1. Unfortunately, ILW's legend of the circular marker is "p_{rep}," which could confuse the reader: It should be understood as a shortcut for "the mean of the sampling distribution of $p_{\mathrm{rep}|d}$."

The same computations can be done for any δ. Some other values are the following: mean$(p_{\mathrm{rep}|d}|\delta = 0) = 0.696$; mean$(p_{\mathrm{rep}|d}|\delta = 0.50) = 0.775$; mean$(p_{\mathrm{rep}|d}|\delta = 2) = 0.995$. The results are illustrated in Figure 2, which corresponds to the first panel of ILW's (2009) Figure 1.

ILW's (2009) simulations are internally consistent, but they ignore the fact that p_{rep} is not intended to estimate a parameter; moreover, they claim that p_{rep} should estimate $\Phi^2(-\delta\sqrt{n/2}) + \Phi^2(\delta\sqrt{n/2})$ (p. 424)—that is, that the sampling mean of $p_{\mathrm{rep}/d}$ for fixed δ should be equal to this quantity (or at least be close to it). Given our considerations above about p_{rep} and $\varphi_{+|\delta}$, it seems very strange that (in our notation) $(\varphi_{+|\delta})^2 + (1-\varphi_{+|\delta})^2$ is considered by ILW to be "the true probability of replication" (and called "Truth" in their Figure 1)—strange, because this parameter clearly does not answer the question, "What is the probability of getting again a positive effect in a replication ($d_{\mathrm{rep}} > 0$)?" This point will be clarified in the next section. The fact that the sampling mean of p_{rep} sensibly differs from ILW's "true probability of replication" can in no way be interpreted as "an indication of bias" (ILW, 2009, p. 424).

ILW's (2009) other criticisms concern the "undesirably high variability of p_{rep}" (p. 425). This high variability applies as well to the sampling distributions of p values (Cumming, 2008) and Bayesian probabilities about δ. Figure 3 shows, for instance, the sampling distribution of $P(\delta{>}0|d)$, the Bayesian posterior probability that δ is positive, assuming the usual noninformative uniform prior. A well-known result is that this probability is equal to $1-p$, where p is the one-sided *p value* associated with the test of the null hypothesis $\delta = 0$ against the alternative $\delta > 0$. The sampling distribution of $P(\delta{>}0|d)$ has a mean that does not estimate any "natural" parameter and has a high variability. Consequently, if you accept ILW's criticism that high vari-

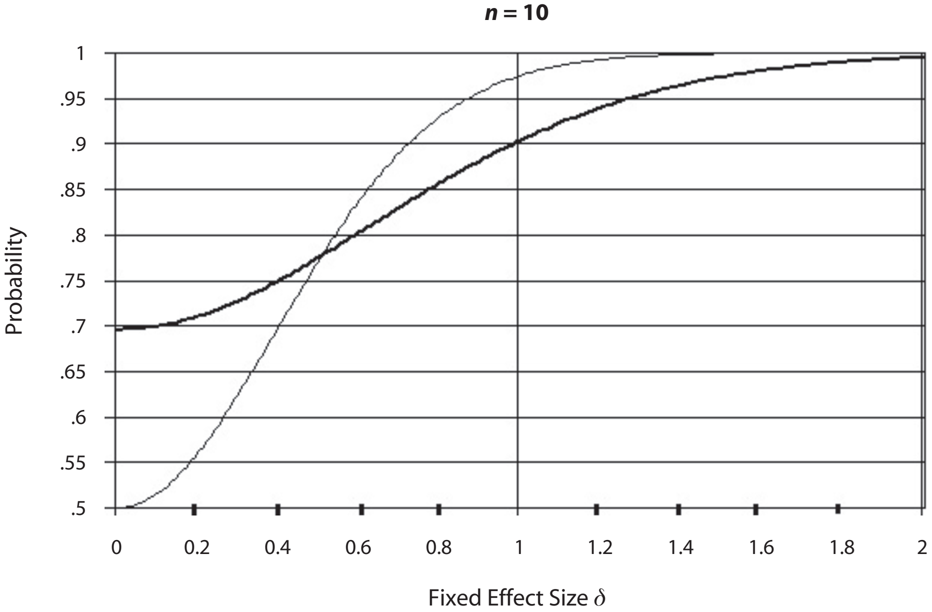

Figure 2. The thick line is the sampling mean of the probability of replication—treated as a random variable $p_{\text{rep}|d}$—as a function of the true effect size δ. It corresponds to the circular markers in the first panel of Iverson, Lee, and Wagenmaker's (2009) Figure 1. The thin line is $(\varphi_{+|\delta})^2 + (1-\varphi_{+|\delta})^2$ (what Iverson et al., 2009, called "Truth"), where $\varphi_{+|\delta}$ is the true probability of observing a positive effect.

ability invalidates the use of a statistic, you should accept its natural consequences and ban any statistical procedure that is designed for a decision process, such as *p* values and Bayesian probabilities. In sum, it is inappropriate to apply the conventional criteria for judging estimators (unbiasedness, minimum variance estimator) to such statistics. But if they are applied, the same brush tars the classic Bayesian and frequentist statistics as well.

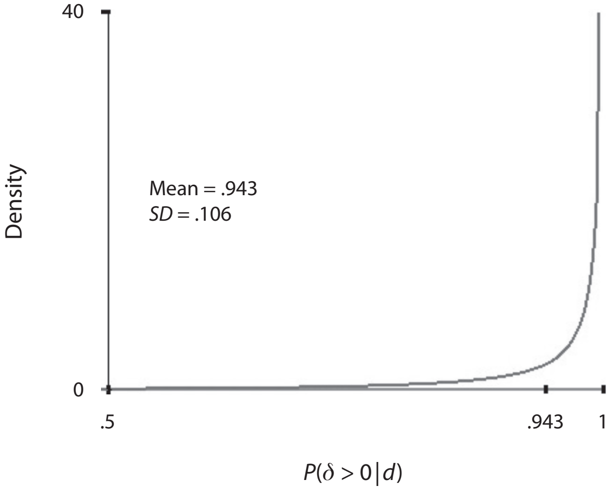

Figure 3. Sampling distribution of the Bayesian posterior probability $P(\delta > 0\,|\,d)$ (assuming a uniform prior) when $n = 10$ and the underlying true effect size is $\delta = 1$. It is also the sampling distribution of $1-p$, where *p* is the one-sided *p value* associated with the test of the null hypothesis $\delta = 0$ against the alternative $\delta > 0$.

Theory: Probability of Replication and Probability of Coincidence

ILW (2009) conflated the probability of replication, p_{rep}, with the probability of *coincidence*. According to ILW, "in his influential paper, Killeen [2005] proposes a measure—the probability of 'replication,' p_{rep}, where replication means 'agreeing in sign.'" This reprise is elliptic: What Killeen actually said was "Define *replication* as *an effect of the same sign as that found in the original experiment*" (p. 346). Manifestly, p_{rep} is a probability about the replicate effect *conditional* on the observed effect ("after observations"). It must not be confused with a *joint* probability, such as "the probability that both *d*, and an imagined replicate observed effect size d_{rep}, have the same sign," which is ILW's definition of p_{rep}. This confusion leads ILW to a misplaced definition of a parameter that they called "true probability of replication." In the following comments, we systematically address the concerns of both frequentists and Bayesians. For a more complete grounding of p_{rep}, turn to Lecoutre, Lecoutre, and Poitevineau (in press).

What Is the Probability of a Replication's Returning an Effect of the Same Sign As That Found in an Original Experiment?

Clearly, this question applies to the situation in which you have collected data that show an effect size of d_{obs}. If you tell a frequentist that you have observed a positive effect ($d_{\text{obs}} > 0$) and ask, "What is the probability of getting again a positive effect in a replication ($d_{\text{rep}} > 0$)?" he or she will say "φ_+ of course: the true sampling probability of observing a positive effect." You are naturally dissatisfied with this answer, since you do not know the value of φ_+; he or she may helpfully add, "I could give you an estimate of φ_+." Perhaps; but that is irrelevant. The question on the table is clearly not about finding a point or interval estimate of φ_+ but, rather, about evaluating, with some inevitable degree of uncertainty, the probability of a particular replication outcome: a predictive inference. As is the case for p values, p_{rep} is not designed for estimating a particular parameter, and it makes little sense to ask whether p_{rep} is a good estimator. It is designed to give the probability that a new experiment will find the same sign of effect.

Frequentists must recognize that p_{rep} is based on a Bayesian (or fiducial) approach, and need to distinguish between two different kinds of probabilities.

1. *Sampling probability (φ_+ known)*. This framing *is equivalent to hypothesizing a fixed value* δ for the true effect size—hence, a hypothesized "true probability of replication" $\varphi_{+|\delta} = \Phi(\delta\sqrt{n/2})$ (see above). In particular, if you hypothesize that δ is 0, then $\varphi_{+|0} = 0.5$, and you know that, in the long run, half the observed effects will be positive and half will be negative. The probability of a same-sign replication is .5, independently of the data in hand (d_{obs}): If you know that the coin is fair, the probability of getting a head on a future toss is exactly .5, independently of the past outcomes. This is summarized in the first column of Table 1.

2. *Posterior predictive probability p_{rep} (φ_+ unknown)*. The framing above is hypothetical: In real situations you have $d_{\text{obs}} > 0$ in hand, but *do not know* the true value of δ. Let us say, for simple illustration, that there exist three kinds of coins: fair coins, two-headed coins, and two-tailed coins. You cannot know what kind of coin is tossed; you know only the outcome. If you have observed one head, what will be the probability of getting another head if the same coin is tossed again? If you hypothesize that a fair coin has been tossed, the answer is .5; if you hypothesize that a two-headed coin was tossed, the answer is 1. (This situation illustrates the fact that it is obviously desirable to use the past outcome to compute this probability, since the first observation ruled out the possibility that it is a two-tailed coin.) But frequentists can go no further.

The Bayesian answer is a *weighted mean* of .5 and 1, the weights being the posterior probabilities of the hypotheses. This is the classical Bayesian notion of *posterior predictive probability*, averaged on the parameter space. In the same way, the probability of a same-sign replication, p_{rep}, is a weighted mean of *all possible* "true probabilities of replications" (see the opening paragraphs). The weights express your uncertainty about the parameter $\varphi_{+|\delta}$—or equivalently here, δ—regarded as random variables, *given the data in hand* (Lecoutre et al., in press). Bayesian predictive probabilities clearly answer the question about evaluating, with some level of uncertainty, the probability of a replication outcome.

ILW's (2009) "True Probability of Replication" Is a "True Probability of Coincidence"

ILW (2009) appear to have recognized that the question is not about estimating $\varphi_{+|\delta}$, but they seem not to have been convinced that p_{rep} is intended to estimate a parameter. With this perspective, they introduced an *arbitrary parameter*, misnamed "the true probability of replication" (p. 424). In their note 1, ILW defined "the true probability of replication for a fixed effect size (i.e., a delta value)" as "the probability an observed effect and its replicate will agree by both having the same sign as δ [plus] the probability they will both agree by having the opposite sign to δ" (p. 428). This verbal definition uses the words "observed" and "replicate" to distinguish between the two effects, but the distinction is illusory, since these two effects (d and d_{rep} in their notations) are actually undistinguished random variables. A more exact description of the probability that they computed is "the sampling probability—conditional on a fixed delta value and before observations—of observing two same-sign effects in two different (but undistinguished, future) independent sets of observations." This is a joint probability about future effects: in our notations, $\varphi^{+2} + (1{-}\varphi^{+})$ (see note 2). This joint probability should not be termed "probability of replication" but, more appropriately, "*probability of coincidence*" (the probability that two future effects will coincide in sign; second column of Table 1). This probability could be used in the situation in which two different investigators plan to run the same experiment. Assuming two identical populations, it is the sampling probability (given a known δ) that the two future experiments will return a same-sign effect. Clearly, estimating the probability of getting two successive heads or two successive tails in the situation where you *know* what kind of coin is tossed (*probability of coincidence*) is profoundly different from evaluating the probability of obtaining a second head in the situation where you *do not know* what kind of coin is tossed (*probability of replication*). Table 1 summarizes the differences between the different kinds of probabilities.

Does p_{rep} actually return accurate predictions? In the mundane world of real data where meta-analyses present numerous accomplished replications, it does (Killeen, 2005). In the world of simulations, it does as well, as is shown in the next section.

Table 1
Three Different Probabilities

Sampling Probabilities δ Fixed		Predictive Probability Averaged on δ
Replication (d_{obs} fixed, d_{rep} random)	ILW's $p_{\text{coincidence}}$ (d and d_{rep} both random)	Killeen's $p_{\text{replication}}$ (d_{obs} fixed, d_{rep} random)
$p([\text{sign}(d_{\text{rep}}) = \text{sign}(d_{\text{obs}})] \mid d_{\text{obs}},\delta)$ $= p(d_{\text{obs}} d_{\text{rep}} > 0 \mid d_{\text{obs}},\delta)$ $= \varphi_{+\|\delta}$ if $d_{\text{obs}} > 0$ and $1{-}\varphi_{+\|\delta}$ if $d_{\text{obs}} < 0$	$p([\text{sign}(d_{\text{rep}}) = \text{sign}(d)] \mid \delta)$ $= p(d d_{\text{rep}} > 0 \mid \delta)$ $= \varphi_{+\|\delta}{}^2 + (1{-}\varphi_{+\|\delta})^2$	$p([\text{sign}(d_{\text{rep}}) = \text{sign}(d_{\text{obs}})] \mid d_{\text{obs}})$ $= p(d_{\text{obs}} d_{\text{rep}} > 0 \mid d_{\text{obs}})$ $= p_{\text{rep}}$

Simulations

Research generally occurs in a world of random effects, where differences in test materials, experimenters, and confederates mean that analysis must cope with variance due not only to sampling error over subjects, but also to that over locales. To bring the analysis into closest relevance to practitioners, the following simulations are of random effects, based on a recent meta-analysis of social science research at large.

Richard, Bond, and Stokes-Zoota (2003) summarized the results of over 25,000 social psychological studies. They converted all effect sizes to $|r|$ and presented them in their Figure 1. Using the r-to-z transform, they are shown at the top of Figure 4. The curve through them is normal with a mean of 0. This symmetry makes sense, since the sign of the effect is conventional (even if it is crucially important to respect in replication attempts). The 75th percentile of that normal distribution falls at an effect size of approximately 0.37; this is, therefore, the median effect size of positive effects, and correspondingly, -0.37 is the median of negative effects. The z-to-r transformation gives an $r = .18$, corresponding to this representative effect size. A value of $r = .18$ is, in fact, the median value of $|r|$ found by Richard and associates. It checks.

Richard et al. (2003) also found that within 18 "literatures" (e.g., aggression, attitudes, attribution . . . social influence), the standard deviations of effect sizes—of the population parameters for that literature across experimental details—were relatively invariant, averaging 0.15 (corresponding to approximately 0.3 in units of d). This variability was found after correcting for the variance attributable to subjects (Hedges & Vevea, 1998). This indicates that any researcher attempting a conceptual replication of prior work (*conceptual* meaning reasonable variation in a procedure that should preserve the main effect) will experience a ceiling on the probability of replicating that effect. In particular, for the typical realization variance found by Richard and associates, it requires an initial effect size of $d_{\text{obs}} = 0.5$ to realistically hope for a 90% replicability (Killeen, 2007). The best the typical experiment ($d_{\text{obs}} = 0.37$) can realistically hope for is 80%, below the threshold of conventional significance—thus, the many failures to replicate (Ioannidis, 2005).

Figure 4 shows the plan of the simulation. (1) On each run, a population parameter δ was sampled from a normal distribution with a mean of 0 and a standard deviation of 0.55, corresponding to the full distribution, the right limb of which was reported by Richard et al. (2003). This determined the "literature" for the run; in Figure 4, it takes the sample value of $\delta \approx 0.75$. (2) The next step within the run was to sample for parameters relevant to the experimental and replication instantiations of the manipulation: the random effects phase. These were sampled from a normal distribution, with a mean of δ and a standard deviation of 0.28, in keeping with the results in Richard et al. The means of these realizations are represented in the figure as δ_1 and δ_2. (3) Then n_{E} samples from the first distribution ($\mu = \delta_1$, $\sigma = 1$) constituted the first experimental group;

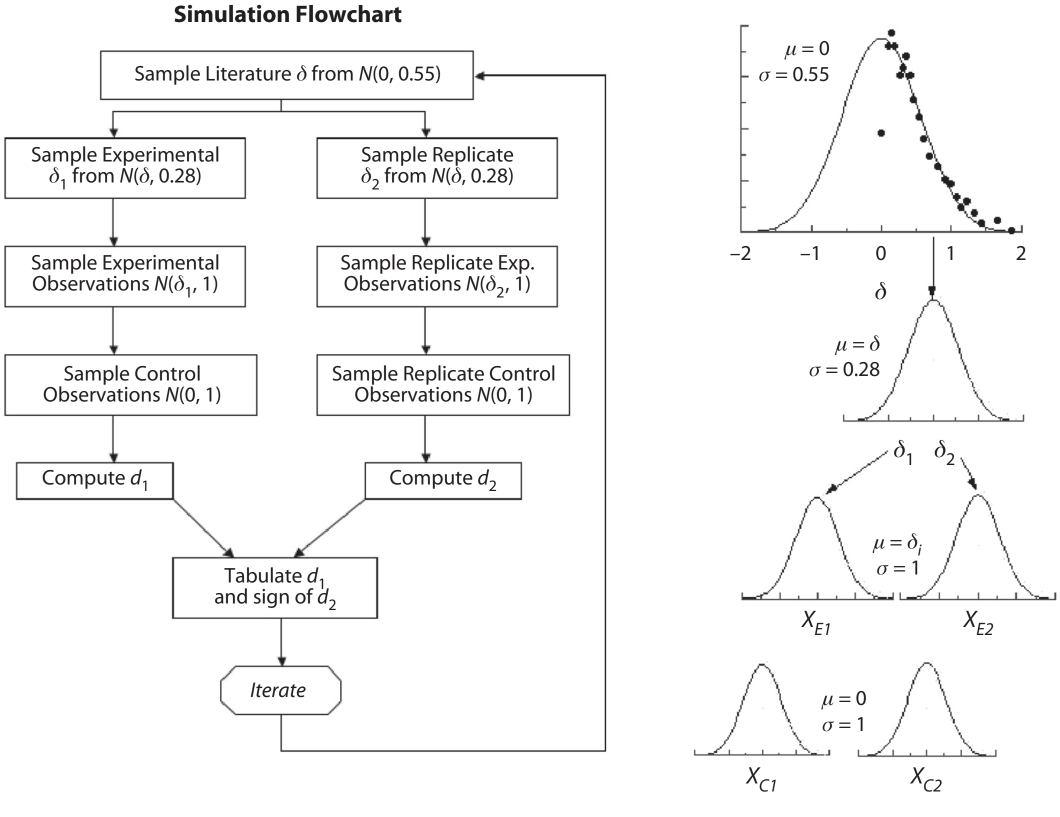
Simulation Flowchart
Sample Literature δ from N(0, 0.55)
Sample Experimental δ1 from N(δ, 0.28)
Sample Replicate δ2 from N(δ, 0.28)
Sample Experimental Observations N(δ1, 1)
Sample Replicate Exp. Observations N(δ2, 1)
Sample Control Observations N(0, 1)
Sample Replicate Control Observations N(0, 1)
Compute d1
Compute d2
Tabulate d1 and sign of d2
Iterate
μ = 0
σ = 0.55
−2
−1
0
1
2
δ
μ = δ
σ = 0.28
δ1
δ2
μ = δi
σ = 1
XE1
XE2
μ = 0
σ = 1
XC1
XC2

Figure 4. Flowchart and exemplary distributions used in the simulations, adapted from Killeen (2007).

a corresponding number of samples, $n_C = n_E$, from a distribution with $\mu = 0$, $\sigma = 1$, constituted the first control sample. (4) The next step was to sample from the same literature for the replication experiment: n_E samples from the second distribution ($\mu = \delta_2$, $\sigma = 1$), constituting the replication experimental group; finally, an equal number of samples $n_C = n_E$ from a distribution with $\mu = 0$, $\sigma = 1$ generated the replication control sample.

(5) At this point, all information about δ, δ_1, and δ_2 was discarded; the effect size measured in the first experiment, d_1, was recorded in a table and, alongside it, whether the replication was a success (same sign) or a failure (different sign). Twenty thousand such runs constituted a simulation for that sample size. (6) The process was repeated for the next sample size.

The results were grouped into nine approximately equally populated bins; the number of successes associated with each bin was divided by the total number of observations in that bin. These constituted the ordinates of Figure 5. For the abscissae, the absolute value of the median effect size within the bin (call it d_{1i}) was selected as representative of the bin and was used to predict the proportion of replications of the same sign (p_{rep}). We computed p_{rep} as $N(d_{1i}, \sigma^2_{d_R})$, where the replication variance is $\sigma^2_{d_R} = 2(\sigma^2_{d_{1i}} + \sigma^2_{\delta_i})$. The variance of effect sizes is $\sigma^2_{d_{1i}} = 4/(n - 4)$, where $n = (n_E + n_C) > 4$, which closely approximates that given by Hedges and Olkin (1985) over the interval $d = \pm 1$, the range within which we work. The random effects realization variance, $\sigma^2_{\delta_i}$, is approximately the same across all literatures and, in the simulation, was therefore kept constant at (0.28) (see note 2). This is the only parametric information carried forward to inform the predictions. It was carried forward because it is something of a universal in social science research. It could be dispensed with by conducting the simulations as a fixed effect model, setting $\sigma^2_{\delta_i}$ to 0. Figure 5 shows that the predictions were accurate over a large range of sample sizes. This should lay to rest the question of accuracy.

In the second half of their article, ILW (2009) conducted simulations superficially similar to these; their results for p_{rep} were discrepant from those for p^*_{rep}. But this is because they compared their p^*_{rep} on the basis of knowledge of δ and the appropriately smaller variance that that entails (their Step 3), with p_{rep} (their Steps 4 and 5) for which knowledge of the parameter is disavowed. Absent knowledge of the parameter, which is the whole point of p_{rep}, the additional step of inferring the posterior distribution (and thence, the prediction) adds that additional source of variance, handicapping it in competition with a better-informed alternative. In predicting replication, if you know the parameter, you should use ILW's p^*_{rep}, which, as we have suggested, is more appropriately thought of as a probability of coincidence (second column of Table 1).

Variability

In the typical experiment, the measure on the experimental group is separated from that on the control group by just over a third of a standard deviation, with a corresponding point–biserial correlation of around $r = .2$ (Richard et al., 2003); our manipulations do not typically control a lot of the variance in our data. Because the original experiment is as subject to sampling error as is a replicate, estimates of replicability are imperfect. With $d_1 = 0.37$ and a total n of 20, as in ILW's (2009) Figure 1, there is a 10% chance that a replication attempt will provide strong evidence *against* the original effect (Killeen, 2005, 2007, shows how to perform the calculation). But this is not a problem for p_{rep} so much as a challenge for our experimental techniques: The same variability is present in other inferential statistics—in particular, p values (Cumming, 2008) and Bayes factors. In their novel deployment of p_{rep}, Ashby and O'Brien (2008) alerted readers to the uncertainty inherent in values of p_{rep} less than .9, a caution we echo. Miller (2009) noted the informational equivalence of many inferential statistics and their generally disheartening performance in predicting replicability of effects and provided the sober counsel with which all writers on this topic will finally agree: "Ultimately, variability must be overcome by increasing sample size and reducing measurement error, not by improving statistical techniques" (p. 634).

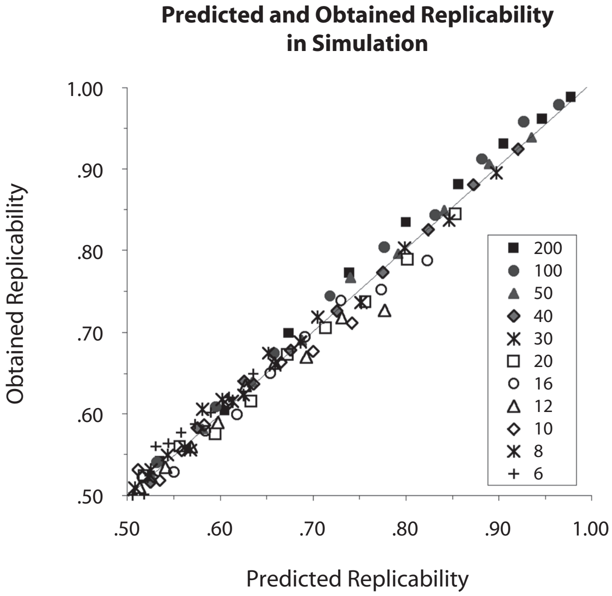

Figure 5. Results of the simulation. Predicted replicability for each run is calculated using the absolute value of the midpoints of nine bins as d_1, and the value of n shown in the legend for that run. The obtained replicability is the proportion of times that d_2 had the same sign as d_1. From "Replication Statistics," by P. R. Killeen, 2007, in J. W. Osborne (Ed.), *Best Practices in Quantitative Methods* (p. 117), Thousand Oaks, CA: Sage. Copyright 2007 by Sage Publications. Adapted with permission.

Conclusion

ILW (2009) concluded that p_{rep} is "a poor estimator," "biased and highly variable." These conclusions follow from mistaking replication with coincidence and from fixing the value of the population parameter, rather than the initial observation. They simulated the mean of the sampling distribution of p_{rep} for fixed true effect size values (varying from 0 to 2) and compared it with the probability of coincidence. No demonstration is needed to state that p_{rep} *does not estimate* this parameter. There is, in fact, no sensible reason for comparing the two quantities. Consequently, the demonstration is misleading, and ILW's conclusions are irrelevant for Killeen's (2005) statistic.

P_{rep} stands on its own as a third way to evaluate data, going from available data to future observations. It combines the standard Bayesian analysis (going from observations to parameters) with the usual frequentist sampling analysis (going from parameters to observations) in the experimentalists' statistical armamentarium. It opens the door to novel applications (Ashby & O'Brien, 2008; Irwin, 2009) and provides opportunities for a decision-theoretic approach to statistical inference (Killeen, 2006).

AUTHOR NOTE

Correspondence concerning this article should be addressed to P. R. Killeen, Department of Psychology, Arizona State University, Box 1104, McAllister St., Tempe, AZ 85287-1104 (e-mail: killeen@asu.edu).

REFERENCES

Ashby, F. G., & O'Brien, J. B. (2008). The p_{rep} statistic as a measure of confidence in model fitting. *Psychonomic Bulletin & Review*, **15**, 16-27. doi:10.3758/PBR.15.1.16

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, **3**, 286-300. doi:10.1111/j.1745-6924.2008.00079.x

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall/CRC.

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. doi:10.1037/1082-989X.3.4.486

Ioannidis, J. P. A. (2005). Why most published research findings are false [Electronic version]. *PLoS Medicine*, **2**. Retrieved August 1, 2008, from http://dx.doi.org/doi:10.1371/journal.pmed.0020124

Irwin, R. J. (2009). Equivalence of the statistics for replicability and area under the ROC curve. *British Journal of Mathematical & Statistical Psychology*, **62**, 485-487. doi:10.1348/000711008X334760

Iverson, G. J., Lee, M. D., & Wagenmakers, E.-J. (2009). p_{rep} misestimates the probability of replication. *Psychonomic Bulletin & Review*, **16**, 424-429. doi:10.3758/PBR.16.2.424

Killeen, P. R. (2005). Replicability, confidence, and priors. *Psychological Science*, **16**, 1009-1012. doi:10.1111/j.1467-9280.2005.01653.x

Killeen, P. R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review*, **13**, 549-562.

Killeen, P. R. (2007). Replication statistics. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 103-124). Thousand Oaks, CA: Sage.

Lecoutre, B., Lecoutre, M.-P., & Poitevineau, J. (in press). Killeen's probability of replication and predictive probabilities: How to compute and use them. *Psychological Methods*.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, **16**, 617-640. doi:10.3758/PBR.16.4.617

Richard, F. D., Bond, C. F. J., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, **7**, 331-363. doi:10.1037/1089-2680.7.4.331

NOTES

[1] This is ILW's (2009) notation. Killeen (2005) differentiated the numbers in experimental and control groups as n_E and n_C, to make it easier to treat cases with different numbers in each. For consistency with the critics, however, we use their notation in this section.

[2] These values agree with ILW's (2009) simulations: "When $n = 10$ and . . . $\delta = 1$ . . . p_{rep} on average gives a value of about .90, . . . , with one standard deviation around the mean extending from about .80 to 1.00" (p. 425).

(Manuscript received October 26, 2008;
revision accepted for publication October 22, 2009.)