p_{rep} Misestimates the Probability of Replication

Geoffrey J. Iverson and Michael D. Lee
Department of Cognitive Sciences
University of California, Irvine

Eric-Jan Wagenmakers
Department of Psychology
University of Amsterdam

The probability of 'replication', p_{rep}, has been proposed as a means of identifying replicable and reliable effects in the psychological sciences. We conduct a basic test of p_{rep} that reveals it misestimates the true probability of replication, especially for small effects. We show how these general problems with p_{rep} play out in practice, when it is applied to predict the replicability of observed effects over a series of experiments. Our results show that, over any plausible series of experiments, the true probabilities of replication will be very different from those predicted by p_{rep}. We discuss some basic problems in the formulation of p_{rep} that are responsible for its poor performance, and conclude that p_{rep} is not a useful statistic for psychological science.

The p_{rep} Measure of Replication

Searching for significant effects in psychological experiments is a risky business, because data are often sparse and noisy. Killeen (2005a) rightly points out that searching for small effects is especially perilous using the contorted logic of null hypothesis significance testing (see Wagenmakers, 2007, for a review). So, in his influential paper, Killeen (2005a; see also Killeen, 2005b, 2005c, 2006; Sanabria & Killeen, 2007) proposes a measure—the probability of 'replication', p_{rep}, where replication means "agreeing in sign"—that is claimed to offer hope.

The simplest way to understand p_{rep} is to consider the standard situation in which data are Normally distributed with a common known variance σ^2, and with an experimental group mean μ_E and control group mean μ_C. If both the experimental and control groups have n subjects, then the observed effect size d is a draw from a Normal distribution with mean $\delta = (\mu_E - \mu_C)/\sigma$, where δ is the 'true' underlying effect size, and variance $2/n$.

Under these assumptions, p_{rep} is derived as the probability that both d, and an imagined replicate observed effect size d_{rep}, have the same sign. A standard Bayesian posterior predictive calculation then gives $p_{\text{rep}} = \Phi\left(|d|\sqrt{n/4}\right)$, as long as a uniform prior is placed on δ (e.g., Doros & Geier, 2005). We give formal details of this derivation in the appendix, but immediately make three clarifying observations.

First, note that it is important to take the absolute value of the effect size in calculating p_{rep}. Otherwise, for example, an observed effect $d = -2$ with $n = 25$ would give a $p_{\text{rep}} < 0.00001$, corresponding to an extremely strong belief that the replicate effect would have a positive sign, which is ridiculous. We mention this point because it is not very clear in the existing p_{rep} literature, where sometimes the absolute value notation has been omitted from key equations.

Secondly, note that our notation differs from Killeen's, who uses n to denote the combined sample size from both the control and experimental groups, whereas we use n for each group separately. We prefer our notation, because it will generalize more naturally to cases where the number of subjects in each group is not the same.

Thirdly, we note that for small sample sizes, Killeen (2005a) promotes the use of an *ad hoc* correction in which n is replaced by $n-2$ (in our notation). This makes a small quantitative difference that disappears quickly as n increases, but does not change the qualitative pattern of our results nor the substantive conclusions at all.

The General Pattern of Misestimation for p_{rep}

In this section we present a general pattern of results that make it clear p_{rep} is a poor estimator. We do this by comparing the true probability of replication for a fixed effect size (i.e., a δ value), to the estimates of the probability of replication provided by p_{rep}.

Each panel of Figure 1 shows, for a different sample size n, a thick dotted line corresponding to the true probability of replication for underlying effect sizes from 0 to 2. This true probability of replication, averaged across all possible sampled observed effects d, is given[1] by $\Phi^2\left(|\delta|\sqrt{n/2}\right) + \Phi^2\left(-|\delta|\sqrt{n/2}\right)$. Each panel in Figure 1 also shows the mean estimates of replication probability provided by p_{rep} for these δ and n values, based on 1,000,000 sampled observed effect sizes. The error bars represent one standard

Address correspondence to: Michael D. Lee, Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, 92697-5100. Telephone: (949) 824 5074. Facsimile: (949) 824 2307. Electronic Mail: mdlee@uci.edu

[1] The first term is the probability an observed effect and its replicate will agree by both having the same sign as δ. The second term is the probability they will both agree by having the opposite sign to δ. See Iverson, Lee, Zhang, and Wagenmakers (submitted) for details.

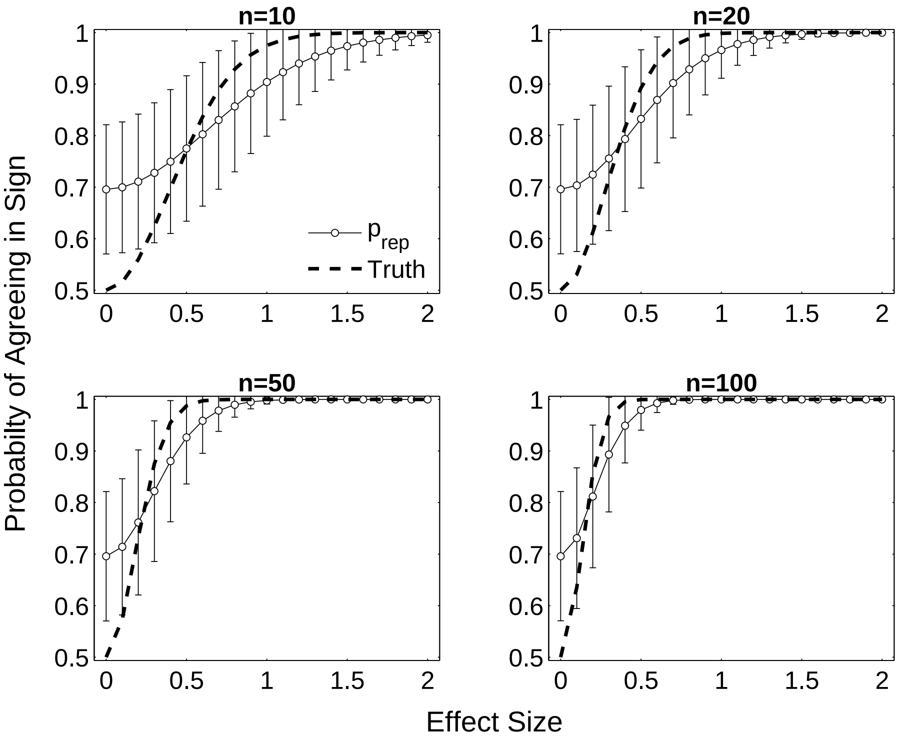

Figure 1. The misestimation of p_{rep}. Each panel shows the true probability of replication (thick broken line) for effect sizes ranging from 0 to 2, the mean performance of p_{rep} (circular markers), and one standard deviation in both directions around the mean (error bars). The four panels correspond to sample sizes of 10, 20, 50 and 100.

deviation above and one standard deviation below the mean values of p_{rep}.

In the language of statistical estimation, the difference between the mean value of p_{rep} and the truth provides an indication of bias, while the size of the error bars provide an indication of variance. For bias, Figure 1 shows that, for small underlying effect sizes, p_{rep} always on average overestimates the probability of replication. For larger effects, p_{rep} then underestimates the true probability of replication, and only becomes well-calibrated for very large effect sizes. The quantitative details of when overestimation becomes underestimation depend on the sample size across the four panels in Figure 1, but the qualitative pattern does not. In fact, every choice of sample size has a curve like those shown, simply shifting further left as sample size increases.

In terms of variability, the error bars in Figure 1 show that p_{rep} is highly variable, except when effect sizes or sample sizes are very large. For example, when $n = 10$ and the underlying true effect size is $\delta = 1$, the actual probability of replication is about 0.95, p_{rep} on average gives a value of about 0.90, but the variability is large, with one standard deviation around the mean extending from about 0.80 to 1.00.

The results in Figure 1 have serious consequences for the performance of p_{rep}. For small effect sizes, where much of the psychological interest lies, and where new experimental findings can make the biggest contribution to the psychological sciences, p_{rep} is highly variable and exaggerates the probability of replication. Only for very large effect sizes does p_{rep} work (approximately) as advertised. Figure 1 suggests that, unless we are willing to believe most experiments have very large effects, p_{rep} will on average lead us to overestimate the probability of replication, and will do so with undesirably high variability. Figure 1 also shows that we cannot safely use p_{rep} to identify replicable or reliable small effects.

The Practical Consequences of Misestimation for p_{rep}

Killeen (2005a, p. 351), in his closing statements, conceives of p_{rep} allowing the management of risk in a research setting:

> “But editors may lower the hurdle for potentially important research that comes with so precise a warning label as p_{rep}. When replicability becomes the criterion, researchers can gauge the risks they face in pursuing a line of study: An assistant professor may choose paradigms in which p_{rep} is typically greater than .8, whereas a tenured risk taker may [pursue] a line of research having p_{rep}s around .6”.

Of course, only clairvoyants can identify those experiments that will give them p_{rep} values of exactly 0.6 or 0.8. This means we cannot simply use the analysis in Figure 1 to look up how p_{rep} will misestimate in practice. While our analysis shows p_{rep} has general problems, it does not make explicit how those problems will play out in practice when p_{rep} is used to make predictions about replicability for observed effect sizes as Killeen (2005a) proposed. In this section we address the problem of misestimation in practice directly.

Research Strategies

Under Killeen's (2005a) risk-management conception, researchers do series of experiments, hunting for replicable and reliable effects, according to some risk management strategy. The more aggressive tenured researchers might choose experiments they believe might have small effects, and avoid doing less interesting experiments where the effect is obvious from the outset. The more conservative untenured researchers might spread their net wider, being happy to do experiments with large underlying effect sizes, but inevitably also doing experiments with small underlying effect sizes.

A sensible way to think about these different risk-seeking profiles is to imagine each attempted experiment having a true but unknown effect size that is drawn from a distribution of possible experiments. The distribution that is used corresponds to the risk management strategy. Four possible strategies are shown in Figure 2. The top panel shows riskier strategies for tenured researchers, focused on small effect sizes. Strategy A assumes the distribution has its mode at zero, while Strategy B makes the more optimistic assumption that researchers are astute enough to be able to place modes on small but genuine effects, and then try to control the variance of their distribution to focus on these effect sizes. Strategies C and D in the bottom panel, for the untenured researcher, follow the same pattern, except now the distributions have greater variance, so that experiments with larger underlying effects are also included in the mix.

All of the strategy distributions are symmetric about zero, because of the nature of effect size measures (i.e., the magnitude of an effect size carries information, but the sign is arbitrary). This symmetry requires, for example, that observed effects of $d = +2$ and $d = -2$ be equally likely for any given strategy. For this reason, it is possible to formulate any strategy more succinctly as a distribution over absolute effect size, in which case the strategies in Figure 2 would become truncated Normal distributions. In these terms, the means for Strategies A and C are zero, and the means for Strategies B and D are 0.2. The standard deviations for Strategies A and B are 0.3, and the standard deviations for Strategies C and D are 0.8.

The Performance of p_{rep}

Whatever strategy researchers use, p_{rep} is supposed to give them the probability that effects they observe for each experiment will be replicated in sign. A p_{rep} value of 0.85 claims there is an 85% probability the next effect will have the same sign, and a 15% chance it will not. It is easy to test the usefulness of p_{rep} as an estimator of these probabilities by simulation. We examined the four strategies shown in Figure 2, and focused on a standard root mean square error (RMSE) measure of the difference between the true probability of replication and the estimate provided by p_{rep}.

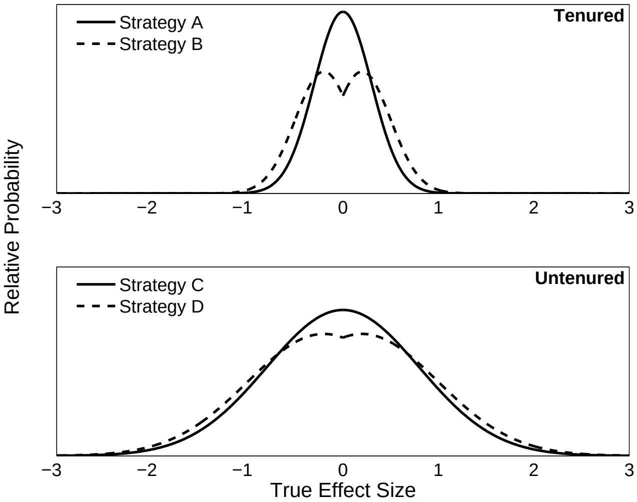

Figure 2. Choosing experiments according to a risk strategy. The top panel shows two possible strategies for focusing on effects with small effect sizes. The bottom panel shows two possible risk strategies that are willing to tackle both small and large effects.

Our simulation test used the following seven steps.

- Step 1: Choose an experiment by sampling from the distribution defined by the risk strategy. Call the true underlying effect size for the particular experiment sampled δ.
- Step 2: Generate the observed effect size from an experiment—which involves experimental and control groups both with n subjects—from the Normal distribution with mean δ and variance $2/n$. Call this d.
- Step 3: Calculate the true probability of replication, which is given by $p^*_{\text{rep}} = \Phi\big(\text{sgn}\,(d)\,\delta/\sqrt{2/n}\big)$.[2]
- Step 4: Calculate $p_{\text{rep}} = \Phi\big(|d|\,\sqrt{n/4}\big)$.
- Step 5: Calculate the Mean Squared Error (MSE) between the true probability of replication, p^*_{rep}, and the estimate p_{rep}. For the tth trial, this is $\text{MSE}_t = \big(p^*_{\text{rep}} - p_{\text{rep}}\big)^2$.
- Step 6: Go back to Step 1 to conduct the next experiment, until a total of T have been completed.
- Step 7: When all T experiments are completed, average the MSEs over all the experiments, and take the square root of this average, to get the final RMSE. That is, calculate $\text{RMSE} = \sqrt{1/T \sum_t \text{MSE}_t}$.

To make the process of the simulation test concrete, the first two example trials from our simulations, using Strategy A with $n = 10$, proceeded as follows. On the first

[2] To calculate this true probability of replication, p^*_{rep}, given a known true effect size δ, sample size n, and an observed effect size d, we are simply finding the area under the Normal distribution with mean δ and variance $2/n$ that has the same sign (i.e., lies on the same side of zero) as d.

Table 1

The RMSE, and the average p_{rep} values (in brackets), for experimental strategies A–D and various sample sizes.

Sample Size	Strategy A		Strategy B		Strategy C		Strategy D	
n=10	0.22	(0.74)	0.22	(0.72)	0.19	(0.82)	0.20	(0.78)
n=20	0.21	(0.79)	0.22	(0.75)	0.16	(0.87)	0.18	(0.83)
n=50	0.18	(0.84)	0.20	(0.81)	0.13	(0.91)	0.15	(0.89)
n=100	0.16	(0.88)	0.17	(0.85)	0.11	(0.94)	0.13	(0.92)

trial δ was sampled to be 0.28, and d was then sampled as 0.90. The true probability of replication is 0.73, and p_{rep} is 0.90, so the mean square error for this experiment is $(0.73-0.90)^2 \approx 0.03$. On the second trial δ was sampled to be -0.51, and d was then sampled as -0.31. The true probability of replication is 0.87, and p_{rep} is 0.67, so the mean square error for this experiment is $(0.87-0.67)^2 \approx 0.04$. The final RMSE measure is the square root of the average of all of the mean square errors calculated in this way.

Table 1 shows the RMSE measures, and the average values of p_{rep}, for Strategies A–D with sample sizes of 10, 20, 50 and 100. These results are based on $T = 1,000,000$ simulated experiments.[3] The RMSE measures can be interpreted as the average 'distance' between the true probability of replication, and the estimate provided by p_{rep}.

It is clear that Table 1 shows p_{rep} is a poor estimator. When risky strategies are in play, or when sample sizes are small, the RMSEs are often over 0.2, which is a very large discrepancy on a probability scale. The situation improves for less risky strategies and larger sample sizes, but even for Strategy C, which regularly does experiments with true effect sizes greater than one, using a sample sizes of 100, so that p_{rep} is giving an average value of 0.94, the RMSE remains above 0.1.

The Underlying Problems

It is well known in statistics that the RMSE measure can be understood as the sum of a bias term and a variance term (e.g., Mood, Graybill, & Boes, 1974). Both need to be low—so that an estimator consistently produces values near the truth—for the RMSE measure to show good performance. So, given the general problems with bias and variance in our first (fixed-effect) analysis in Figure 1, it is not surprising to find the poor RMSE values in our second (random-effect) analysis in Table 1.

Why does p_{rep} have these basic problems? We tackle bias first. Although authors have tried to derive and interpret p_{rep} from a number of statistical perspectives, the most useful one (as usual) is the Bayesian perspective, alluded to in our introduction, and detailed in the appendix. There it is shown that p_{rep} can be derived from the posterior predictive distribution $d_{\text{rep}} \mid d$ when an improper flat prior is placed on the true effect size δ (e.g., Doros & Geier, 2005).

The problem is that the assumption of a flat prior on the true underlying effect size is not a good one. Nobody believes Nature makes effect sizes of 500, 50, 5, and 0.5 all equally likely experimental outcomes, yet this is exactly what p_{rep} assumes. It is true that Bayesian statisticians often use flat priors to express ignorance about the value of location parameters that lie on arbitrary scales. But effect sizes, by design, come normalized on an invariant and readily interpreted scale, for which there is strong and important prior information. For this reason, no Bayesian would argue for a flat prior on effect sizes. Indeed, in Bayesian statistics, it is standard practice to use prior distributions that put more mass on small effect sizes than on large effect sizes (e.g., Gönen, Johnson, Lu, & Westfall, 2005; Lee, 2008; Zellner & Siow, 1980).

Discussing the Bayesian interpretation of p_{rep}, Sanabria and Killeen (2007) defend the choice of the improper flat prior, arguing that it has the advantage of not "including information not specifically contained within the experiment itself" (p. 481). This does not seem to us to be a very convincing argument, because the scale of an effect size is part of understanding an experiment. If we tell you we have just collected a response time, you do not know whether we are going to say 0.3, 3, 30, or any other number, because you do not know whether our measurements are in seconds, tenths of a second, hundredths of a second, or any other scale. Here a flat uniform prior appropriately captures what you know, which is almost nothing, and lets the data speak for themselves. But if we tell you we have just collected an effect size, you know we are much more likely to say 0.3 than 3, and that we are not going to say 30. The experiment, by virtue of the normalized effect size scale, contains specific prior information that must be part of our inferences. The flat uniform prior assumed by p_{rep} is a strange distortion of what is known, and forces the data to say things they do not mean, and are not true.

Intuitively, because p_{rep} assumes a flat prior, it does not give sufficient prior probability to small effect sizes around zero. This means it is overly optimistic about the magnitude of the effect sizes it expects to exist in Nature, and so overestimates the probability of replication based on the single observed effect size it receives as a datum. It follows that it is possible to improve the bias p_{rep} by making more realistic prior assumptions. The generalization of p_{rep} given by p^{θ}_{rep} in Equation 5 of the appendix allows for non-uniform

[3] Using the *ad hoc* correction in which n is replaced by $n-2$ (in our notation), some of the entries in Table 1 change in the second decimal place by 0.01, but most do not change at all.

priors, taking the form of zero-mean Normal distributions over effect sizes. Setting this prior to Strategy A or C, which are also zero-mean Normal distributions over effect sizes, removes the bias from p_{rep}. It also, however, results in smaller p_{rep} values.

Removing the variability of p_{rep} is more challenging. One basic problem is the conception—which is fundamental to the definition of p_{rep}—that agreeing in sign is a good way to measure replication. This conception forces p_{rep} to attempt the estimation of a binary quantity, which makes high variability almost inevitable. Removing the bias in p_{rep}, using the approach outlined above, will not address the problem of variability. Note that in Table 1, even for Strategy C with $n = 100$, where the effect sizes and sample size are large enough that bias is not severe, p_{rep} continues to perform poorly in terms of its RMSE, because it remains variable.

Conclusion

We have presented direct tests of p_{rep} as an estimate of the probability replication for different underlying effect sizes, and as a predictor of replication for different experimental strategies. p_{rep} performs poorly on both because it is biased and highly variable. It overestimates the probability of replication for small observed effect sizes, which are exactly those it was developed to help diagnose, and are the most important ones for helping develop models and theories in psychology. The nature of the bias that leads p_{rep} to exaggerate evidence is general, and is based, at least in part, on a basic mis-assumption about the information conveyed by effect sizes. In addition p_{rep} suffers from having to estimate the binary quantity of agreement in sign, which makes it a highly variable measure.

In short, our results show that p_{rep} is a poor estimator and predictor. Thus, while we agree with several of the motivations behind p_{rep}—including the focus on prediction to evaluate models and data, and the need to change current statistical practices in psychology—we do not view p_{rep} itself as a suitable alternative.

Acknowledgments

We thank Peter Killeen, Jeff Rouder, and two anonymous reviewers for useful feedback, and especially thank Rob Nosofsky for detailed comments.

References

Doros, G., & Geier, A. B. (2005). Probability of replication revisited: Comment on “an alternative to null–hypothesis significance tests”. *Psychological Science*, *16*, 1005–1006.

Gönen, M., Johnson, W., Lu, Y., & Westfall, P. (2005). The Bayesian two-sample t-test. *The American Statistician*, *59*(3), 252–257.

Iverson, G. J., Lee, M., Zhang, S., & Wagenmakers, E.-J. (submitted). p_{rep}: An agony in five fits.

Killeen, P. R. (2005a). An alternative to null–hypothesis significance tests. *Psychological Science*, *16*, 345–353.

Killeen, P. R. (2005b). Replicability, confidence, and priors. *Psychological Science*, *16*, 1009–1012.

Killeen, P. R. (2005c). Tea tests. *The General Psychologist*, *40*(2), 12–15.

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

Lee, M. D. (2008). BayesSDT: Software for Bayesian inference with signal detection theory. *Behavior Research Methods*, *40*(2), 450–456.

Mood, A., Graybill, D., & Boes, D. (1974). *Introduction to the theory of statistics* (Third Edition ed.). New York: McGraw Hill.

Sanabria, F., & Killeen, P. R. (2007). Better statistics for better decisions: Rejecting null hypotheses statistical tests in favor of replication statistics. *Psychology in the Schools*, *44*(5), 471–481.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 585–603). Valencia: University Press.

Appendix: Formal Details

This appendix gives formal details of the derivation of p_{rep} in a Bayesian context, and closely follows the original work of Doros and Geier (2005). The prior on the true effect size δ we assume to be Normal (we will later let its variance go to infinity so that the Normal becomes the uniform assumed by p_{rep}), with variance τ^2

$$\delta \sim \text{Normal}\big(0, \tau^2\big). \tag{1}$$

For experimental and control groups with equal numbers of subjects, n, the observed effect size is then

$$d \sim \text{Normal}\big(\delta, \frac{2}{n}\big). \tag{2}$$

The posterior of the true effect size conditional on this observation is now

$$\delta \mid d \sim \text{Normal}\big(d\theta, \frac{2}{n}\theta\big), \tag{3}$$

where

$$\theta = \frac{\frac{n}{2}\tau^2}{1 + \frac{n}{2}\tau^2}.$$

This makes the posterior predictive density of d_{rep}, the next effect

$$d_{\text{rep}} \mid d \sim \text{Normal}\big(d\theta, \frac{2}{n} + \frac{2}{n}\theta\big). \tag{4}$$

Since p_{rep} is the probability that d and d_{rep} agree in sign, it is given by

$$p^{\theta}_{\text{rep}} = \Pr\big(dd_{\text{rep}} \geq 0 \mid d\big) = \Phi\left(\frac{|d|\,\theta\sqrt{\frac{n}{2}}}{\sqrt{1+\theta}}\right). \tag{5}$$

Now we let the Normal become uniform, by letting $\tau \to \infty$, so that $\theta \to 1$, and we get

$$p_{\text{rep}} = \Phi\left(|d|\sqrt{\frac{n}{4}}\right), \tag{6}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal random variable.