Psychonomic Bulletin & Review
2010, 17 (2), 270-272
doi:10.3758/PBR.17.2.270

The random effects p_{rep} continues to mispredict the probability of replication

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

and

Eric-Jan Wagenmakers
University of Amsterdam, Amsterdam, The Netherlands

In their reply, Lecoutre and Killeen (2010) argue for a random effects version of p_{rep}, in which the observed effect from one experiment is used to predict the probability that an effect from a different but related experiment will have the same sign. They present a figure giving the impression that this version of p_{rep} accurately predicts the probability of replication. We show that their results are incorrect and conceptually limited, even when corrected. We then present a meaningful evaluation of the random effects p_{rep} as a predictor and find that, as with the fixed effects p_{rep}, it performs very poorly.

This reply addresses the two issues raised by Lecoutre and Killeen (2010; hereafter, LK). The first is their claim that we conflated two probabilities. The second is their claim that p_{rep} is an accurate predictor.

The first issue is easy to address. LK (2010) assert that Iverson, Lee, and Wagenmakers (2009) conflated two probabilities: the probability of coincidence and Killeen's (2005) probability of replication. On the basis of this supposed conflation, LK argue that "ILW's conclusions are irrelevant for Killeen's (2005) statistic" (p. 269). The fact of the matter is otherwise. We did not confuse these two probabilities. In Iverson, Lee, and Wagenmakers—and all of our earlier commentaries (Iverson, Lee, Zhang, & Wagenmakers, 2009; Iverson, Wagenmakers, & Lee, in press)—we used exactly the fixed effects p_{rep} definition that appears in the third column of Table 1 in LK. We most certainly did not confuse the statistic p_{rep} with the parameter p_{coinc} (for probability of coincidence), and we invite readers to verify this for themselves.

The second claim regarding the accuracy of p_{rep} is a more important source of disagreement. In their reply, LK (2010) stress a p_{rep} that is conceptually different from the fixed effects version, which they claim returns accurate predictions for both simulated and real-world data (LK, 2010, p. 266). Both versions of p_{rep} use a known effect size from an experiment. In the fixed effects formulation, $p_{\text{rep}}^{\text{F}}$, the goal is to predict the probability that a replication of the same experiment would yield an effect size of the same sign as the original. In the random effects version, $p_{\text{rep}}^{\text{R}}$, the goal is to use an effect size from one experiment to predict the probability of getting an effect of the same sign from a *different* experiment, albeit one coming from the same literature. This new formulation seems to us a strange goal for empirical science. Does it make sense to think that, having observed people preferring oval to square faces, we want to predict whether they will prefer natural to morphed faces?

But whatever the conceptual challenges, it is possible to continue analyzing $p_{\text{rep}}^{\text{R}}$ as a statistic. In more or less technical terms, our previous commentaries showed that $p_{\text{rep}}^{\text{F}}$ made poor predictions about the true replication probability. This reply extends those analyses to evaluate $p_{\text{rep}}^{\text{R}}$.

The Meaning of LK's (2010) Figure 5

The flowchart simulation presented by LK (2010), culminating in their Figure 5, gives the illusion of successful prediction under uncertainty. The abscissa is $p_{\text{rep}}^{\text{R}}$. The ordinate is a different random effects formulation of p_{rep}, for which we derive an analytic expression,[1] and which we denote $p_{\text{rep}}^{\text{O}}$. LK use numerical simulation to evaluate this ordinate.

The relationship between the functions $p_{\text{rep}}^{\text{R}}$ and $p_{\text{rep}}^{\text{O}}$, for the same set of total sample sizes N as that considered by LK (2010), is shown in our Figure 1A. Each line corresponds to a different sample size, and, by choosing different effect sizes, the whole curve relating the two p_{rep} versions can be traced out. We were surprised that these patterns did not seem to agree with Figure 5 in LK, and so we used their flowchart to calculate the results numerically. Using the same binning, median summaries, and other display assumptions that they adopted, our results are shown in Figure 1B and match our analytic results in Figure 1A. After some experimentation, we found that we could approximate LK's Figure 5 by doubling the sample size when generating the per-group simulated effect sizes that computationally approximate $p_{\text{rep}}^{\text{O}}$, but not changing the sample size when calculating $p_{\text{rep}}^{\text{R}}$. The results of this flawed simulation are shown in Figure 1C and seem to match LK's Figure 5. On this basis, we speculate that LK's simulations might have been confused by the different use in this debate of the same symbol n to denote either the total number of subjects (which we denote N here) or the number of subjects per group (which we denote n).

The main problem with LK's (2010) Figure 5, however, is not that it was incorrectly computed, but that it is conceptually limited and potentially misleading. Contrary to the labeling of the ordinate, their figure does not compare $p_{\text{rep}}^{\text{R}}$ with the true probability of replication. A sensible evaluation of $p_{\text{rep}}^{\text{R}}$ must involve a comparison with the true probability of replication, which we call p_{rep}^{*}. In the next section, we present a more complete evaluation of $p_{\text{rep}}^{\text{R}}$, in

M. D. Lee, mdlee@uci.edu

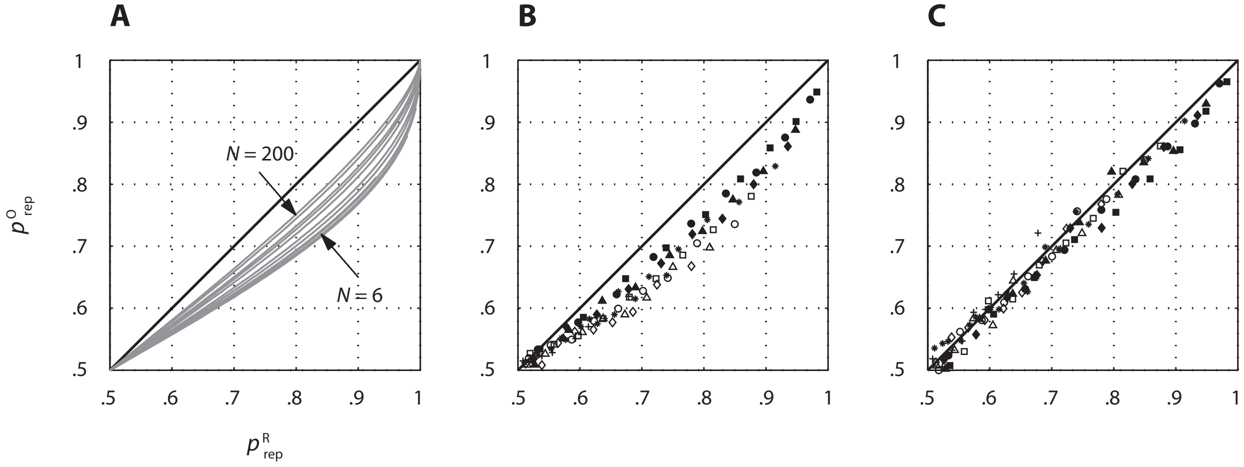

Figure 1. (A) The analytic relationship between p^{R}_{rep} and p^{O}_{rep}. (B) Simulation using Lecoutre and Killeen's (2010) flowchart confirming the analytic relationship. (C) Flawed simulation confusing sample sizes, producing a result that appears to match that in Figure 5 in Lecoutre and Killeen.

which we explicitly compute p^{*}_{rep} and compare it with the prediction made by p^{R}_{rep}.

Evaluation of p^{R}_{rep}

To quantify the performance of p^{R}_{rep}, we repeated our earlier evaluation, using the standard root-mean square error of prediction (RMSEP) measure of performance, now using exactly the random effects environment defined by LK (2010). Our simulation test uses the same approach as that in Iverson, Lee, and Wagenmakers (2009), with the inclusion of Steps 1a and 1b to deal with the change in framework for p^{R}_{rep}, and a change in Step 4 to calculate p^{R}_{rep} itself. More specifically, we performed the following steps.

Step 1. Choose a literature by sampling from the distribution defined by LK, with a mean of 0 and a standard deviation of $\tau = 0.55$. Call the effect size sampled δ_0.

Step 1a. Choose a first experiment by sampling from the distribution defined by LK, with a mean of δ_0 and a standard deviation of $\tau' = 0.28$. Call the effect size sampled δ_1.

Step 1b. Choose a second experiment by sampling from the distribution defined by LK, with a mean of δ_0 and a standard deviation of $\tau' = 0.28$. Call the effect size sampled δ_2.

Step 2. Generate the observed effect size from an experiment—which involves experimental and control groups,

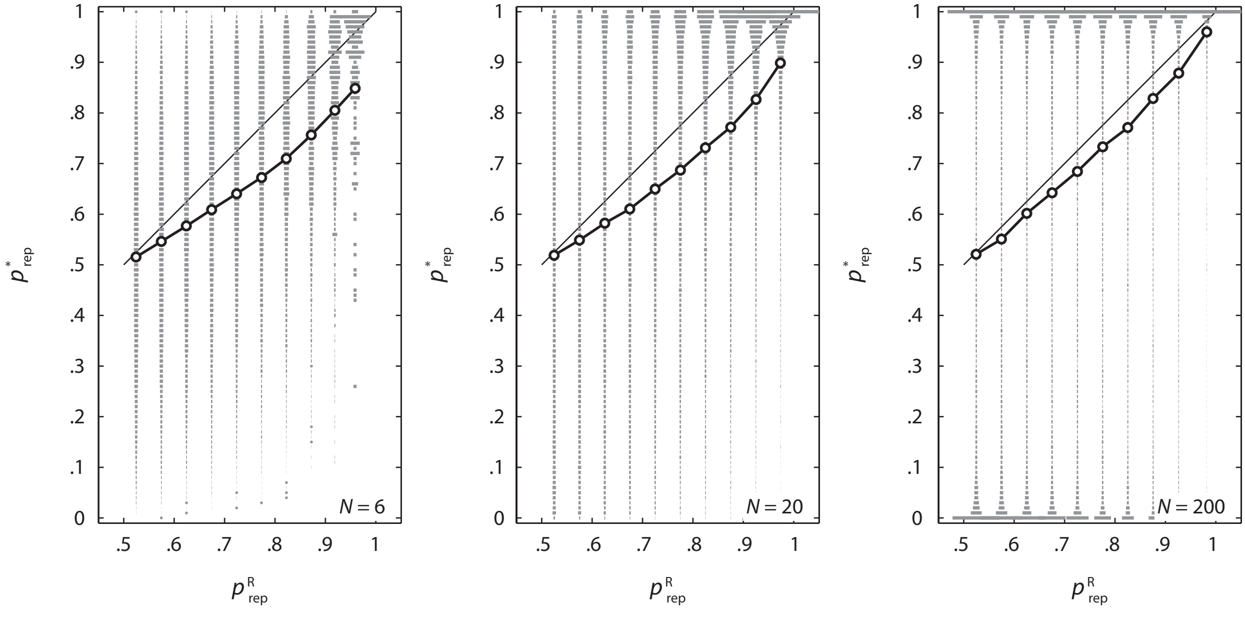

Figure 2. The relationship between p^{R}_{rep} and p^{*}_{rep}, for $N = 6$ (left panel), $N = 20$ (middle panel), and $N = 200$ (right panel). For each p^{R}_{rep} bin, the distribution of p^{*}_{rep} is shown by the width of the gray bars. The pattern of change in the mean of p^{*}_{rep} is shown by the line with circular markers and can be compared with the thin solid line denoting perfect performance.

both with n subjects—from the normal distribution with a mean of δ_1 and a variance of $2/n$. Call this d.

Step 3. Generate many candidate effect sizes, d_{rep}, for the second experiment from the normal distribution with a mean of δ_2 and a variance of $2/n$. Use these effect sizes to find the true probability of replication, by calculating the proportion that agree in sign with d. We used 100,000 d_{rep} values, which is enough to agree closely with the analytic result. This true proportion is p^*_{rep} and is what $p^{\text{R}}_{\text{rep}}$ is trying to predict.

Step 4. Following the definition given by LK, calculate

$$p^{\text{R}}_{\text{rep}} = \Phi\left(\frac{|d|/\sqrt{2}}{\sqrt{\frac{2}{n} + \tau'^2}} \right),$$

where $\tau' = 0.28$.

Step 5. Calculate the mean squared error of prediction (SEP) between the true probability of replication, p^*_{rep}, and $p^{\text{R}}_{\text{rep}}$. For the tth trial, this is $\text{SEP}_t = (p^*_{\text{rep}} - p_{\text{rep}})^2_t$. If $p^{\text{R}}_{\text{rep}}$ is doing its job as a predictor, it should be close to the true probability of replication, and the SEP should be small.

Step 6. Go back to Step 1 to conduct the next experiment, until a total of T experiments have been completed.

Step 7. When all T experiments are completed, average the SEPs over all the experiments and take the square root of this average, to get the final RMSEP. That is, calculate

$$\text{RMSEP} = \sqrt{1 / T \sum_t \text{SEP}_t}\,.$$

Figure 2 shows the relationship between $p^{\text{R}}_{\text{rep}}$ and p^*_{rep} for three choices of N. It does this by defining a series of bins for $p^{\text{R}}_{\text{rep}}$ and drawing the density of p^*_{rep} for each bin, using gray bars. Also shown, by circular markers, is the average of p^*_{rep} in each bin. These averages correspond to $p^{\text{O}}_{\text{rep}}$. It is clear from Figure 2 that p^*_{rep} is almost always highly variable and that the average is typically biased, with $p^{\text{R}}_{\text{rep}}$ overstating $p^{\text{O}}_{\text{rep}}$. For example, with $N = 20$, $p^{\text{R}}_{\text{rep}} \approx .92$ overstates the average true $p^*_{\text{rep}} \approx .83$.

Table 1 reports the RMSEP measures of discrepancy between what $p^{\text{R}}_{\text{rep}}$ predicts and what it ought to predict. The RMSEP measure is shown for the same total sample sizes and effect size bins as those considered by LK (2010), based on $T = 100{,}000$ experiments for each sample size. The RMSEPs, which combine the bias and variance visually evident in Figure 2, clearly show the poor performance of $p^{\text{R}}_{\text{rep}}$ as a predictor. What it predicts is very often .3, .4, or .5 from what should have been predicted, which is a very large difference on the probability scale 0 to 1.

Table 1
The Root-Mean Square Error of Prediction for $p^{\text{R}}_{\text{rep}}$, for Each Combination of Effect Size Bins, d_{bin}, and Total Sample Sizes $N = 2n$ Considered by Lecoutre and Killeen (2010)

	Total Sample Size (N)										
d_{bin}	6	8	10	12	16	20	30	40	50	100	200
0.04	.46	.45	.44	.43	.41	.41	.40	.38	.38	.32	.30
0.12	.45	.46	.45	.45	.44	.43	.40	.40	.38	.35	.31
0.20	.46	.47	.46	.46	.44	.44	.43	.41	.40	.36	.33
0.29	.47	.47	.47	.46	.46	.45	.44	.42	.40	.36	.33
0.38	.48	.48	.48	.48	.47	.47	.44	.43	.41	.37	.33
0.48	.48	.49	.49	.48	.48	.48	.45	.42	.41	.37	.33
0.59	.49	.50	.50	.50	.49	.48	.45	.42	.41	.35	.30
0.73	.50	.51	.51	.50	.50	.48	.45	.41	.39	.32	.26
0.96	.51	.53	.53	.51	.49	.47	.42	.38	.35	.27	.21

Conclusion

We think our earlier analyses of $p^{\text{F}}_{\text{rep}}$ showed that *Psychological Science* was right to reverse its earlier recommendation and to remove mention of p_{rep} from its instructions to authors. We think our analyses of $p^{\text{R}}_{\text{rep}}$ should similarly discourage its use.

AUTHOR NOTE

Correspondence concerning this article should be addressed to M. D. Lee, Department of Cognitive Sciences, University of California, Irvine, CA, 92697-5100 (e-mail: mdlee@uci.edu).

Note—The Editor limited this invited reply to a word count of about 1,000 words.

REFERENCES

IVERSON, G. J., LEE, M. D., & WAGENMAKERS, E. (2009). p_{rep} misestimates the probability of replication. *Psychonomic Bulletin & Review*, **16**, 424-429.

IVERSON, G. J., LEE, M. D., ZHANG, S., & WAGENMAKERS, E. (2009). p_{rep}: An agony in five fits. *Journal of Mathematical Psychology*, **53**, 195-202.

IVERSON, G. J., WAGENMAKERS, E., & LEE, M. D. (in press). A model averaging approach to replication: The case of p_{rep}. *Psychological Methods.*

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

LECOUTRE, B., & KILLEEN, P. R. (2010). Replication is not coincidence: Reply to Iverson, Lee, and Wagenmakers (2009). *Psychonomic Bulletin & Review*, **17**, 263-269.

NOTE

[1] Available at www.socsci.uci.edu/~mdlee/IversonEtAl2010_Note.pdf

(Manuscript received November 18, 2009;
revision accepted for publication January 12, 2010.)