The Random-Effects p_{rep} Continues to Mispredict the Probability of Replication: Online Appendix

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

Eric-Jan Wagenmakers
Department of Psychology
University of Amsterdam

Abstract

This document is the technical on-line appendix to "The Random-Effects p_{rep} Continues to Mispredict the Probability of Replication".

Background

Killeen (2007) and Lecoutre and Killeen (this volume) describe a numerical simulation for the calculation of p_{rep} within a random effects model. In the language that is used to describe the model, one first selects a "literature", as a draw from a prior distribution $\delta_0 \sim N\big(0, \tau^2\big)$. Given a literature, "within-literature" parameters δ_1 and δ_2 are drawn independently from a population distributed as $N\big(\delta_0, \tau'^2\big)$. Finally, values of effect statistics d and d_{rep} are drawn independently from populations distributed respectively as $N\big(\delta_1, 2/n\big)$ and $N\big(\delta_2, 2/n\big)$; each of the two-group experiments that generate d and d_{rep} is based on the same per-group sample size of $n = N/2$, so that N is the total sample size for an experiment. We follow Lecoutre and Killeen (this volume) in assuming that the variance governing the data from each experiment is the same, and is known.

The variances τ^2 and τ'^2 are estimated by Killeen (2007) from a meta-analysis of effect magnitudes obtained across many published social science literatures, In the following we assume $\tau^2 = 0.3025$ ($\tau = 0.55$) and $\tau'^2 = 0.0784$ ($\tau' = 0.28$), as quoted by Killeen (2007).

In his original article Killeen gave an explicit analytical expression for p_{rep} in a special case of the random-effects model. Tailored to the assumption of known data

variance this expression reads

$$p_{\text{rep}}^{R} = \Phi\left(\frac{|d|}{\sqrt{\frac{4}{n} + 2\tau'^2}}\right).$$

Notation. We shall often find it convenient to write in terms of $z = d\sqrt{n/2}$, $\omega^2 = \tau^2 n/2$, and $\omega'^2 = \tau'^2 n/2$. For example, the above equation becomes

$$p_{\text{rep}}^{R} = \Phi\left(\frac{|z|/\sqrt{2}}{\sqrt{1+\omega'^2}}\right). \tag{1}$$

Note that this expression comes about by assuming that the prior governing δ_0 is the improper uniform density (Killeen, 2005). Note too that Equation 1 degenerates as $\tau'^2 \to 0$ to the expression

$$p_{\text{rep}} = \Phi\left(\frac{|z|}{\sqrt{2}}\right) \tag{2}$$

that is familiar from a fixed-effects model.

Example. The numerical value of p_{rep}, for given experimental data, thus depends on the model chosen to compute it. For instance, suppose an experiment based on $n = 25$ yields $d = 0.56$. From Equation 2 we obtain $p_{\text{rep}} = 0.92$, whereas Equation 1 gives the less optimistic value 0.84, reflecting the additional "realization variance" τ'^2.

What is p_{rep} Trying to Predict?

Despite our misgivings about the term "replication" (see Iverson, Lee, Zhang, & Wagenmakers, 2009), we shall retain it. But to understand the construction of p_{rep} in a random-effects model only underscores our misgivings. Having observed experimental data (d, n) an experimenter is asked to predict the probability that a second independent experiment within the same literature (but addressing a possibly quite different research question) will yield data (d_{rep}, n) for which $dd_{\text{rep}} \geq 0$, and we are puzzled why this particular event deserves the term *replication*. At any rate, the experimenter is asked to predict the value of

$$p_{\text{rep}}^{*} = \Pr\left(dd_{\text{rep}} \geq 0 \mid \delta_2, d\right) = \Phi\left(\delta_2 \ \text{sgn}d\sqrt{n/2}\right). \tag{3}$$

For this difficult task Killeen (2005) proposed the posterior predictive probability

$$p_{\text{rep}} = E\left(p_{\text{rep}}^{*} \mid d\right) = \int p_{\text{rep}}^{*} f\left(\delta_2 \mid d\right) \ \mathrm{d}\delta_2 = \Pr\left(dd_{\text{rep}} \geq 0 \mid d\right). \tag{4}$$

In Killeen (2007) and again in Lecoutre and Killeen (this volume) it is claimed that p_{rep} is quite accurate as a predictor of p_{rep}^{*} and Figure 6 in Killeen (2007) and the same Figure 4 in Lecoutre and Killeen (this volume) appear to bear out this claim.

But the claim flies in the face of commonsense. The only thing that our experimenter can know about d_{rep} resides in the data (d, n). Certainly those data can be used profitably to update distributions of prior parameters δ_0 and δ_2 to their posterior counterparts. All the same, considerable uncertainty attends those posterior parameters, and in turn that uncertainty combines with sampling variability in d_{rep} to produce large variability in the posterior predictive distribution $f\left(d_{\text{rep}} \mid d\right)$ on which the calculation of p_{rep} is based. We give detailed calculations, both numerical and analytical, to support our view.

Evaluating p_{rep} as a Predictor: Our Way

If p_{rep} is to be useful as a predictor of p^*_{rep} it will necessarily have to possess a small mean-squared error of prediction, MSEP:

$$\text{MSEP} = \iint \left(p_{\text{rep}} - p^*_{\text{rep}}\right)^2 f\left(\delta_2, d\right) \, \mathrm{d}\delta_2 \, \mathrm{d}d. \tag{5}$$

However, numerical calculations show that MSEP is large, in agreement with commonsense but quite at odds with the simulations of Killeen (2007) and Lecoutre and Killeen (this volume). For typical numerical values of $\sqrt{\text{MSEP}} = \text{RMSEP}$ see Table 1 of our reply.

Evaluating p_{rep} as a Predictor: The Way of Lecoutre and Killeen

Killeen (2007) and Lecoutre and Killeen (this volume) plot what they call "obtained" p_{rep} against what they call "predicted" p_{rep}. There is no mystery as to "predicted" p_{rep}: it is p^R_{rep} as calculated by Equation 1. "Obtained" p^O_{rep} requires a new calculation based on the random-effects model that we described in detail above. Let us now carry out this calculation.

We have, by definition, for the "obtained" version

$$p^O_{\text{rep}} = \int p^*_{\text{rep}} f\left(\delta_2 \mid d\right) \, \mathrm{d}\delta_2. \tag{6}$$

The posterior density $f\left(\delta_2 \mid d\right)$ can be computed from

$$f\left(\delta_2 \mid d\right) = \int f\left(\delta_2 \mid \delta_0\right) f\left(\delta_0 \mid d\right) \, \mathrm{d}\delta. \tag{7}$$

once the posterior density $f\left(\delta_0 \mid d\right)$ is calculated. This last task is not difficult, and one finds that

$$\delta_0 \mid d \sim N\left(d\frac{\omega^2}{1+\omega^2+\omega'^2}, \frac{2}{n}\frac{\omega^2\left(1+\omega'^2\right)}{1+\omega^2+\omega'^2}\right). \tag{8}$$

It follows from Equations 7, 8 and the model assumption $\delta_2 \mid \delta_0 \sim N\left(\delta_0, \tau'^2\right)$ that

$$\delta_2 \mid d \sim N\left(d\frac{\omega^2}{1+\omega^2+\omega'^2}, \frac{2}{n}\left(\omega'^2 + \frac{\omega^2\left(1+\omega'^2\right)}{1+\omega^2+\omega'^2}\right)\right). \tag{9}$$

Finally, integrating out δ_2 one has

$$d_{\text{rep}} \mid d \sim N\left(d\frac{\omega^2}{1+\omega^2+\omega'^2}, \frac{2}{n}\left(\left(1+\omega'^2\right)\left(1+\frac{\omega^2}{1+\omega^2+\omega'^2}\right)\right)\right) \tag{10}$$

and Equation 6 is explicitly evaluated from Equation 10 as the "obtained"

$$p^O_{\text{rep}} = \Phi\left(\frac{|z|\left(\frac{\omega^2}{1+\omega^2+\omega'^2}\right)}{\sqrt{\left(1+\omega'^2\right)\left(1+\frac{\omega^2}{1+\omega^2+\omega'^2}\right)}}\right). \tag{11}$$

We thus come to understand that Kileen's (2007) Figure 6, and Figure 4 in Lecoutre and Killeen (this volume), is merely a plot of one version of p_{rep}, which we have denoted p^O_{rep} for which $\tau^2 = 0.3025$ in Equation 11, against another version, p^R_{rep} for which $\tau^2 = \infty$ in Equation 11 to give Equation 1. The functional dependence of "obtained" p^O_{rep}, given by Equation 11 on "predicted" p^R_{rep}, given by Equation 1 is shown in Figure **??** of the main body of our rejoinder. This very strange plot in no way directly addresses, as we do in terms of RMSEP, the *performance of* p^R_{rep} *as a predictor for* p^*_{rep}. The different calculations of p_{rep}, one by Equation 11, the other by Equation 1, introduces a model-dependent "bias" that shows up in a decomposition of MSEP that we discuss later.

Example (Continued). With $d = 0.56$, $n = 25$, $\tau^2 = 0.3025$, and $\tau'^2 = 0.0784$, Equation 11 gives $p_{\text{rep}} = 0.76$. Killeen (2005) and Lecoutre and Killeen (this volume) are clear about "the uncertainty inherent in values of p_{rep} less than 0.9." In view of this acknowledged "uncertainty", which Lecoutre and Killeen (this volume) do not attempt to quantify, let us compute how large an initial effect must be so that (with the above values of n and of variances τ^2 and τ'^2) one is assured that $p_{\text{rep}} \geq 0.9$. This is a straightforward calculation from Equation 11 and we discover that $|z| \geq 3.547$; equivalently, for a per-group sample size of $n = 25$, we require $|d| \geq 1$. In other words, only initial effects that most of us would agree are *obviously* real qualify as reliably replicable. One does not need to compute p_{rep} to be confident that such large experimental effects will likely be found by others. But as things stand in the literature, p_{rep} is recommended for use with much smaller observed effects, and it is for those smaller effects that we have found the use of p_{rep} to be most problematic.

Credible Intervals for p^*_{rep}

In Iverson et al. (2009), within the context of a fixed-effects model, we wondered why a Bayesian would report a single number p_{rep}, the expected value of p^*_{rep} conditional on d, rather than report the full posterior distribution of p^*_{rep} expressed as a function of $\delta_2 \mid d$. In particular, it is straightforward to compute credible intervals for p^*_{rep} and we do so now for our canonical example. We consider four combinations of values of the prior variances τ^2 and τ'^2:

(a) $\tau^2 = \infty$ and $\tau'^2 = 0$
(b) $\tau^2 = 0.3025$ and $\tau'^2 = 0$

These two combinations each correspond to a fixed effects model. The combination (a) gives p_{rep} as described by Killeen (2005) early on in his original article; i.e., p_{rep} is computed from Equation 2. The combination (b) gives an example of a fixed effects model calculation for p_{rep} that was discussed at some length in Iverson, Wagenmakers, and Lee (in press), where it was denoted p^{θ}_{rep} and shown to take the value $\Phi\left(|z|\,\theta/\sqrt{1+\theta}\right)$ where $\theta = \omega^2/(1+\omega^2)$.

(c) $\tau^2 = \infty$ and $\tau'^2 = 0.0784$
(d) $\tau^2 = 0.3025$ and $\tau'^2 = 0.0784$.

These last two combinations require random effects model calculations. Combination (c) corresponds to placing a flat, improper prior on δ_0. The final combination (d) reflects the model proposed on the basis of meta-analytic considerations that were discussed above.

For each of these combinations we now give a 95% credible interval of values for p^*_{rep} conditional on our canonical data $d = 0.56$, $n = 25$. Each interval is accompanied by the corresponding value of p_{rep}, the conditional expected value of p^*_{rep}. Because the density of p^*_{rep} is in each case markedly skewed to the left our 95% credible intervals are all one-sided.

(a) $p_{\text{rep}} = 0.92$, interval $[.63, 1)$
(b) $p_{\text{rep}} = 0.88$, interval $[.54, 1)$
(c) $p_{\text{rep}} = 0.84$, interval $[.20, 1)$
(d) $p_{\text{rep}} = 0.76$, interval $[.12, 1)$.

In all cases the credible intervals are very broad and reflect considerable uncertainty about values of p^*_{rep}. The accompanying values of p_{rep} simply do not capture this uncertainty. Note that the random-effects models involve greater uncertainty than their fixed-effects counterparts. This is easy to understand from our analysis above, and is why RMSEP values are even larger for random-effects models than for their fixed-effects counterparts.

MSEP Decomposed

From Equation 5 we have

$$\begin{aligned}
\text{MSEP} &= \iint \left(p_{\text{rep}} - p^*_{\text{rep}}\right)^2 f\left(\delta_2, d\right)\ \mathrm{d}\delta_2\ \mathrm{d}d \\
&= \int \left[\int \left(p_{\text{rep}} - p^*_{\text{rep}}\right)^2 f\left(\delta_2 \mid d\right)\ \mathrm{d}\delta_2\right] f\left(d\right) \mathrm{d}d \\
&= \int \left[\text{Var}\left(p^*_{\text{rep}} \mid d\right) + \left(E\left(p^*_{\text{rep}} \mid d\right) - p_{\text{rep}}\right)^2\right] f\left(d\right) \mathrm{d}d. \qquad (12)
\end{aligned}$$

This decomposition of MSEP shows that the uncertainty in p^*_{rep}, as given by its posterior variance, provides the dominant contribution to MSEP. This fact is yet another expression of our analysis. The second term in square brackets in Equation 12 involves a model-dependent bias which provides a typically much smaller contribution than the variance. This bias is, in effect, what Lecoutre and Killeen (this volume) discovered in their simulations and promptly confused with accuracy of prediction.

How p_{coinc} Arises as an Average Target for p_{rep}

A standard way to evaluate an estimator is to study its mean-squared-error performance. A similar measure can be employed for problems of prediction (as here). In ILW we considered the squared-error of prediction $\left(p_{\text{rep}} - p^*_{\text{rep}}\right)^2 \mid \delta, d$ averaged over many pairs (δ, d) according to various joint densities $f\left(\delta, d\right) = f\left(d \mid \delta\right) f\left(\delta\right)$ that correspond to the various research strategies we envisaged. The standard fixed-effects model is assumed whence it follows that Killeen's $p_{\text{rep}} = \Phi\left(|z|/\sqrt{2}\right)$ and $p^*_{\text{rep}} = \Pr\left(dd_{\text{rep}} \geq 0 \mid \delta, d\right) = \Phi\left(\Delta\ \text{sgn}d\right)$ is the true probability of replication in whose evaluation we have used the convenient abbreviation $\Delta = \delta\sqrt{n/2}$. Steps 1 and 2 of our simulation algorithm provide the draws (δ, d). The remaining steps 3–7 provide the appropriate mean-squared-error of prediction:

$$\text{MSEP} = \iint \left(p_{\text{rep}} - p^*_{\text{rep}}\right)^2 f\left(\delta_2, d\right)\ \mathrm{d}\delta_2\ \mathrm{d}d$$

It is useful to decompose this double integral by an initial integration over d, followed by integration over δ. We have

$$\begin{aligned}
\text{MSEP} &= \int \left[\int \left(p_{\text{rep}} - p^*_{\text{rep}}\right)^2 \mathrm{d}d\right] f\left(\delta\right) \mathrm{d}\delta \\
&= \int \left[\text{Var}\left(p^*_{\text{rep}} \mid \delta\right) + \left(E\left(p^*_{\text{rep}} \mid \delta\right) - p_{\text{rep}}\right)^2\right] f\left(\delta\right) \mathrm{d}\delta.
\end{aligned}$$

Note that

$$\begin{aligned}
E\left(\left(p_{\text{rep}} - p^*_{\text{rep}}\right) \mid \delta\right) &= E\left(p_{\text{rep}} \mid \delta\right) - E\left(p^*_{\text{rep}} \mid \delta\right) \\
&= E\left(p_{\text{rep}} \mid \delta\right) - \left(\Phi^2\left(\Delta\right) + \Phi^2\left(-\Delta\right)\right) \\
&= E\left(p_{\text{rep}} \mid \delta\right) - p_{\text{coinc}}
\end{aligned}$$

is a form of bias, the extent to which $E\left(p_{\text{rep}} \mid \delta\right)$ differs from its expected target value $E\left(p^*_{\text{rep}} \mid \delta\right) = p_{\text{coinc}} = \Pr\left(dd_{\text{rep}} \geq 0 \mid \delta\right)$. We also have

$$\text{Var}\left(p_{\text{rep}} - p^*_{\text{rep}} \mid \delta\right) = \text{Var}\left(p_{\text{rep}} \mid \delta\right) + \text{Var}\left(p^*_{\text{rep}} \mid \delta\right) - 2\text{Cov}\left(p_{\text{rep}}, p^*_{\text{rep}} \mid \delta\right),$$

from which we obtain the following decomposition:

$$\text{MSEP} = \int \left[\text{MSEE}_\delta + \text{Var}\left(p^*_{\text{rep}} \mid \delta\right) - 2\text{Cov}\left(p_{\text{rep}}, p^*_{\text{rep}} \mid \delta\right)\right] f\left(\delta\right) \text{d}\ \delta, \tag{13}$$

in which we have

$$\text{MSEE}_\delta = \text{Var}\left(p_{\text{rep}} \mid \delta\right) + \left(E\left(p_{\text{rep}} \mid \delta\right) - p_{\text{coinc}}\right)^2 = E\left(\left(p_{\text{rep}} - p_{\text{coinc}}\right)^2 \mid \delta\right).$$

For small values of the non-centrality parameter $\Delta = \delta\sqrt{n/2}$, common enough in our science, the first term MSEE dominates the contribution of the other two terms. The acronym MSEE stands for mean-squared-error of estimation and is computed *as if* p_{rep} was employed solely for the purpose of estimating $E\left(p^*_{\text{rep}} \mid \delta\right) = p_{\text{coinc}}$. The decomposition in Equation 13 of MSEP shows that MSEE is critical to an understanding of the ability of p_{rep} to predict p^*_{rep} and justifies plotting p_{coinc} as an "average target" for the predictions of p_{rep}, as in Figure 1 in ILW and Figure 6 in ILZW. The lessons learned from the decomposition in Equation 13 stand in sharp distinction to the view of LK who claim that "ILWs conclusions are irrelevant for Killeens statistic.". Readers can decide for themselves whose view of matters is the more compelling, informative, and relevant.

References

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.

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., & Killeen, P. R. (this volume). Replication is not coincidence: Reply to Iverson, Lee and Wagenmakers (2009). *Psychonomic Bulletin & Review*.