

## $p_{\text{rep}}$ : An agony in five Fits

Geoffrey J. Iverson<sup>a,b,\*</sup>, Michael D. Lee<sup>a,b</sup>, Shunan Zhang<sup>a,b</sup>, Eric-Jan Wagenmakers<sup>c</sup>

<sup>a</sup> Department of Cognitive Sciences, University of California, Irvine, United States

<sup>b</sup> Institute for Mathematical and Behavioral Sciences, University of California, Irvine, United States

<sup>c</sup> Department of Psychology, University of Amsterdam, Netherlands

### ARTICLE INFO

#### Article history:

Received 3 June 2008

Available online 22 November 2008

#### Keywords:

$p_{\text{rep}}$

Probability of replication

Posterior prediction

### ABSTRACT

In 2005 *Psychological Science*, the flagship journal of the Association for Psychological Science, began their current practice of asking contributors to compute the statistic  $p_{\text{rep}}$  in lieu of the traditional  $p$ -value. In a polemic comprising five Fits we argue that  $p_{\text{rep}}$  is misnamed, commonly miscalculated, misapplied outside a narrow scope, and its large variability often produces values that invite mistrust and mislead the interpretation of data.

Published by Elsevier Inc.

### Prelude to the Agony

“Come, listen, my men, while I tell you again,  
The five unmistakable marks  
By which you may know, wheresoever you go,  
The warranted genuine Snarks.”

The Hunting of the Snark: FIT THE SECOND, The Bellmans Speech. Lewis Carroll, 1876.

In the May 2005 issue of *Psychological Science* Peter Killeen introduced the statistic  $p_{\text{rep}}$  to the psychological community. He describes  $p_{\text{rep}}$  as follows:

“The statistic  $p_{\text{rep}}$  estimates the probability of replicating an effect. It 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,  $p_{\text{rep}}$  provides all of the information now used in evaluating research, while avoiding many of the pitfalls of traditional statistical inference”. (Killeen, 2005a, Abstract).

At the time James Cutting was chief editor of *Psychological Science* and in an Acknowledgment (Cutting, 2005) that appeared in the December 2005 issue of *Psychological Science*, he wrote “and the General Article by Peter Killeen in the May issue may change how all psychologists report their statistics”. This prediction has turned out to be accurate. Currently, about 60% of contributors to

*Psychological Science* submit values of  $p_{\text{rep}}$  when reporting their statistical analyses.

$p_{\text{rep}}$  is intended to be read “probability of replication”, and gives the very strong impression that experiments yielding large values of  $p_{\text{rep}}$  (currently *Psychological Science* regards  $p_{\text{rep}} \geq 0.85$  as large<sup>1</sup>) are replicable with high probability. Recently the euphemisms ‘reliable’ and ‘robust’ have crept into use, so that, for example,  $p_{\text{rep}} = 0.92$  is said to indicate a reliable experimental finding. Whatever term is used, the unfortunate and misleading impression is that  $p_{\text{rep}} = 0.92$  indicates an experimental effect has been established. This impression does not encourage substantive replication. If an experimental effect is remotely plausible and  $p_{\text{rep}} = 0.92$ , why bother to replicate?

For its calculation,  $p_{\text{rep}}$  requires an analytical context, and to keep matters as simple as possible we shall assume throughout that this context is provided by the independent groups design in which the same number of measurements  $n$  is provided by each of an ‘experimental’ and a ‘control’ group.<sup>2</sup> All measurements are assumed to be mutually independent and normally distributed,

<sup>1</sup> There is no editorial statement that stamps  $p_{\text{rep}} \geq 0.85$  as the gold standard. Indeed, Killeen (2005a,b,c) suggested  $p_{\text{rep}} \geq 0.90$ . However, authors publishing in *Psychological Science* routinely declare values of  $p_{\text{rep}} = 0.86$  and above as signaling significant effects. The first clear signs of hesitation occur when  $p_{\text{rep}} = 0.85$ , with some authors happy to declare this value significant, whereas others are reluctant to do so.

<sup>2</sup> Note that Killeen 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 approach, because it generalizes more naturally to cases where the number of subjects in each group is not the same.

\* Corresponding address: Department of Cognitive Sciences, 3151 Social Sciences Plaza, University of California, Irvine, CA 92697-5100, United States.

E-mail addresses: [giverson@uci.edu](mailto:giverson@uci.edu) (G.J. Iverson), [mdlee@uci.edu](mailto:mdlee@uci.edu) (M.D. Lee), [szhang@uci.edu](mailto:szhang@uci.edu) (S. Zhang), [ej.wagenmakers@gmail.com](mailto:ej.wagenmakers@gmail.com) (E.-J. Wagenmakers).

with a common known<sup>3</sup> variance  $\sigma^2$ . The parameter of interest to the experimenter is the population effect  $\delta = (\mu_E - \mu_C) / \sigma$  and is estimated by the experimental or *substantive* effect  $d = (\bar{x}_E - \bar{x}_C) / \sigma$ . Clearly  $d \sim N(\delta, \frac{2}{n})$  and, as is familiar from elementary statistical theory,  $d$  is ‘best unbiased’ for  $\delta$ . The related quantity  $z = d\sqrt{\frac{n}{2}}$  is a familiar test statistic in this context. Under the standard null hypothesis  $H_0 : \delta = 0$ ,  $z$  is distributed as a standard normal variate (mean 0, variance 1) and one rejects  $H_0$  whenever  $|z| \geq z_{\alpha/2}$  in carrying out the level- $\alpha$  Neyman–Pearson test procedure. Equally familiar is the practice of reporting an associated *probability value*, or *p-value* for short; *p-values* attach themselves to test statistics and in the present context the (two-sided) *p-value* attached to the statistic  $|z|$  is given by

$$p\text{-value} = 2\Phi\left(-|d|\sqrt{\frac{n}{2}}\right) = 2\Phi(-|z|). \quad (1)$$

Killeen (2005a,b,c) rejects much of the standard frequentist estimation and inference machinery. He has no time for estimation:

“But it is rare for psychologists to need estimates of parameters . . .” (Killeen, 2005a, p. 345);

and even less for frequentist inference:

“Our unfortunate historical commitment to significance tests forces us to rephrase [these] good questions in the negative, attempt to reject those nullities, and be left with nothing we can logically say about the questions—whether  $p = .100$  or  $p = .001$ ” (Killeen, 2005a, pp. 345–346).

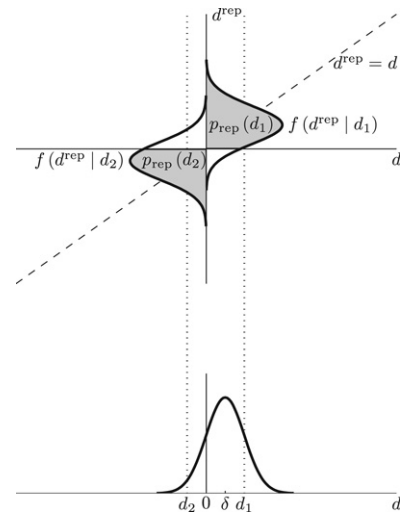
Of course, Killeen is not alone in harboring a critical view of frequentist inference. We hold similar opinions. He is also not alone in calling for an alternative methodology. Now the Bayesian School has elaborated a principled, coherent and readily interpretable alternative to classical estimation and inference.

Killeen declares that his alternative is not Bayesian (Killeen, 2005a). Indeed, he offers his ideas as an alternative that avoids the “Bayesian dilemma” (of having to specify a prior distribution on  $\delta$ ). But as we shall soon see,  $p_{\text{rep}}$  is a Bayesian calculation, though one that is not carried out on a routine basis in Bayesian inference.

Killeen and *Psychological Science* propose that experimenters report an (estimate of) the probability that a repetition  $d^{\text{rep}}$  of an existing experimental effect  $d$  will agree with  $d$  in direction, and to do so in lieu of a conventional *p-value*. This *probability of replication*  $p_{\text{rep}}$  seems new, exciting, and extremely useful. Despite appearances however  $p_{\text{rep}}$  is *misnamed*, commonly *miscalculated* even by its progenitors, *misapplied* outside a common but otherwise very narrow scope, and its seductively large values can be seriously *misleading*. In short, *Psychological Science* has bet on the wrong horse, and nothing but mischief will follow from its continued promotion of  $p_{\text{rep}}$  as a scientifically informative predictive probability of replicability.

**FIT THE FIRST:** In which  $p_{\text{rep}}$  is misnamed

“When I use a word”, Humpty Dumpty said, in a rather scornful tone, “it means just what I choose it to mean—neither more nor less.” *Through the Looking-Glass: Humpty Dumpty*, Lewis Carroll, 1872.



**Fig. 1.** Two independent experimental effects  $d_1$  and  $d_2$  are drawn from the distribution  $f(d | \delta)$  generating the data. Each draw gives rise to a different value of  $p_{\text{rep}}$ , shown by shaded areas. Note that if the true state of nature  $\delta$  is close enough to zero,  $d_1$  and  $d_2$  can have opposite signs. Even so, it is clear that  $p_{\text{rep}}$  is always greater than 0.5.

Killeen (2005a) chooses to “Define replication as an effect of the same sign as that found in the original experiment” (p. 346, emphasis in original). We think this definition is unfortunate and belies normal usage of the terms ‘replicate’ and ‘reliable’.

To attach a probability to this definition requires a model, and despite the obvious “Bayesian dilemma” Killeen invokes two Bayesian models, the fixed effects model and the random effects model. In the fixed effects model independent experiments are literally replicas of one another. That is, they are identical in all respects save for sampling variability, and that variability is the only source of differences among experimental outcomes. Let us call this model  $M_1$  to distinguish it from the random effects model  $M_2$  in which independent repetitions of an experimental protocol combine uncertainty in the population effect parameter  $\delta$  with sampling variability. The standard calculation of  $p_{\text{rep}}$  is carried out under model  $M_1$ :

$$p_{\text{rep}} = \Pr(d \text{ and } d^{\text{rep}} \text{ agree in sign} | d, M_1).$$

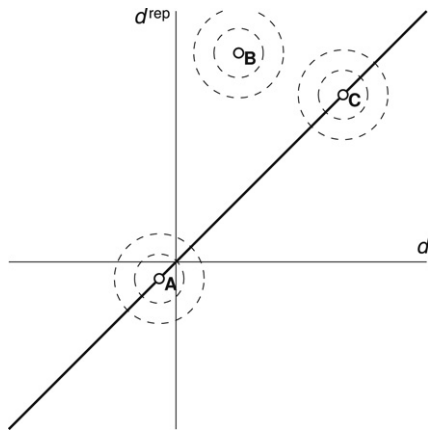
The calculation of  $p_{\text{rep}}$  is pictured in Fig. 1. It is the larger of the areas subtended by the posterior predictive  $f(d^{\text{rep}} | d)$  above and below zero. Since  $f(d^{\text{rep}} | d)$  is not available in frequentist theory,  $p_{\text{rep}}$  is a Bayesian construct.

We take exception to the terminology and notation that attends the definition of  $p_{\text{rep}}$ . The following definition seems more in line with standard English dictionaries and with dictionaries of statistical terms.

**Definition 1.** Independent experimental effects  $d_1$  and  $d_2$  replicate if (and only if) they are each generated under model  $M_1$ . That is, if they are each generated by the same value of  $\delta$ .

Many experimental designs involve comparisons that invite checks of no effect (e.g., no expected effect of order of treatment or of sex or of age cohort). It is anticipated that these comparisons will rarely be significant, and at the same time it is expected that others repeating the same comparisons would reach similar conclusions. That is, experimental comparisons that everyone expects to reflect no or at most a very small effect are nonetheless thought of as highly replicable. This circumstance, which is a commonplace in every empirical science, is entirely in line with the above definition. In such cases measured effects will, over replications, bounce about zero, and there will be a low probability, near 50%, that any two randomly chosen effects agree in sign. For  $p_{\text{rep}}$  however, which

<sup>3</sup> This unrealistic assumption is one of convenience only. It can be dropped, but to do so would involve us in analytical complications that distract from our main purpose. Our critique of  $p_{\text{rep}}$  in no way depends on the assumption of known variance.



**Fig. 2.** The distinction between the notions of ‘replication’ and ‘concurrence’, illustrated by three combinations of  $d$  and  $d^{\text{rep}}$ . The points A, B and C show different states of nature. The circular contours around each indicate the joint distribution of  $d$  and  $d^{\text{rep}}$  in each case. Combination A replicates but does not necessarily concur. Combination B concurs but does not replicate. Combination C replicates and concurs.

places a premium on experimental effects agreeing in sign, these reliable and replicable null experimental outcomes (which seem so essential for the construction of uncluttered and workable theory), are deemed unlikely to replicate and are scorned as unreliable.

To put things another way: if experimental effects are truly generated under model  $M_1$ , they will necessarily replicate according to our definition and it is then most puzzling why one goes to the trouble of computing the probability  $1 - p_{\text{rep}}$  that they will not. Likewise, if repetitions of an experiment are generated under the random effects model  $M_2$  then, according to our definition, they (almost certainly) will not replicate, so why ought one compute the probability  $p_{\text{rep}}$  that they will?

**Definition 2.** Two real numbers  $x_1$  and  $x_2$  *concur* if they agree in sign. That is,  $x_1$  and  $x_2$  concur if  $x_1 x_2 \geq 0$ .

We believe that  $p_{\text{rep}}$  is *misnamed*:  $p_{\text{rep}} = \Pr(d \text{ and } d^{\text{rep}} \text{ concur} \mid d) = \Pr(dd^{\text{rep}} \geq 0 \mid d)$ , and a more appropriate notation would employ  $p_{\text{concur}}$  in place of  $p_{\text{rep}}$ . We shall nonetheless retain the notation  $p_{\text{rep}}$  throughout.

The distinction between replication and concurrence is shown pictorially in Fig. 2, in terms of three different combinations of  $d$  and  $d^{\text{rep}}$ . For true states of nature  $\delta$  falling on the heavy diagonal line, effects  $d$  and  $d^{\text{rep}}$  replicate by definition. This means the combination of parameters A shows that observed effects can replicate but do not always concur. Conversely, combination B shows that observed effects can concur but not replicate. Only for the combination C do  $d$  and  $d^{\text{rep}}$  both replicate and concur.

**FIT THE SECOND:** In which  $p_{\text{rep}}$  is miscalculated

“Two added to one—if that could be done,  
It said, “with one’s fingers and thumbs!”,  
Recollecting with tears how, in earlier years  
It had taken no pains with its sums.

The Hunting of the Snark: FIT THE FIFTH, The Beaver’s Lesson. Lewis Carroll, 1876

Of the 60% or so of authors who currently report  $p_{\text{rep}}$  values in *Psychological Science*, a large majority use the recipe of Killeen (2005c)<sup>4</sup>:

“In particular, whenever a  $p$  value has been calculated, one can immediately infer  $p_{\text{rep}}$  by (a) calculating the  $z$ -score corresponding to  $1 - p$ , (b) dividing by the square root of 2, and (c) finding the probability associated with this new  $z$ -score:

$$p_{\text{rep}} = \Phi \left[ \left( \Phi^{-1} [1 - p] / \sqrt{2} \right) \right]. \quad (2)$$

Unfortunately, the computations of authors following this recipe are often wrong. The standard analytical expression for  $p_{\text{rep}}$  is<sup>5</sup>

$$p_{\text{rep}} = \Phi \left( |d| \sqrt{\frac{n}{4}} \right). \quad (3)$$

Here  $d$  is, as defined above, the observed effect in a comparison of two independent groups, each involving samples of size  $n$ . The accompanying (two-sided)  $p$ -value is given in Eq. (1).

Putting Eqs. (1) and (3) together gives

$$p_{\text{rep}} = \Phi \left[ \frac{\Phi^{-1} \left( 1 - \frac{p}{2} \right)}{\sqrt{2}} \right]. \quad (4)$$

The difference between Eqs. (2) and (4) appears to be minor. The  $p$ -value in Eq. (2) is not halved as it is in Eq. (4) but otherwise the two formulas are identical. Of course the two formulas Eq. (2) and Eq. (4) give different numerical results – a calculation via Eq. (2) is always smaller than via Eq. (4) – but often these differences are rather modest.

In its information for contributors, *Psychological Science* gives the following examples<sup>6</sup>:

“Thus, typical statistical reports would follow formats like these:  
 $t(50) = 2.68$ ,  $p_{\text{rep}} = .95$ ,  $d = 0.76$ ;  $F(1, 30) = 4.69$ ,  
 $p_{\text{rep}} = .90$ ,  $\eta^2 = .135$ ; or  $\beta = .61$ ,  $p_{\text{rep}} = .99$ ,  $d = 1.56$ ”.

For the first two examples, the correct calculation of  $p_{\text{rep}}$  via Eq. (4) gives, in turn,  $p_{\text{rep}} = .97$  and  $p_{\text{rep}} = .91$ . These values are sufficiently close to the ones quoted in the Journal, namely  $p_{\text{rep}} = .95$  and  $p_{\text{rep}} = .90$ , to elicit little more than a shrug. All the same there is unnecessary confusion over how to compute  $p_{\text{rep}}$  from a given  $p$ -value and it seems to us worthwhile to clarify the matter.

It might be argued that Eq. (2) is appropriate to the  $p$ -value from testing a one-sided hypothesis, and in part this is true. Since the one-sided  $p$ -value is one-half of the two-sided  $p$ -value based on the same data, Eqs. (2) and (4) should yield the same numerical answer. To see how things can (and presently do) go awry, consider how the editors of *Psychological Science* obtained  $p_{\text{rep}} = .95$  from the fact that  $t(50) = 2.68$ . This value of Student’s  $t$  statistic gives  $p = .01$  (two-sided) and  $p = .005$  (one-sided). From Eq. (4) or (2) we have (correctly)

$$p_{\text{rep}} = \Phi \left[ \frac{\Phi^{-1} (.995)}{\sqrt{2}} \right] = \Phi \left[ \frac{2.58}{\sqrt{2}} \right] = \Phi [1.824] = .966.$$

What *Psychological Science* appears to have done instead is to compute the two-sided  $p$ -value,  $p = .01$ , and to plug that value into the formula Eq. (2) appropriate to the one-sided  $p$ -value. That mistaken calculation gives

$$p_{\text{rep}} = \Phi \left[ \frac{\Phi^{-1} (.99)}{\sqrt{2}} \right] = \Phi \left[ \frac{2.33}{\sqrt{2}} \right] = \Phi [1.648] = .95.$$

<sup>5</sup> An explicit calculation is indicated below in Eq. (7).

<sup>6</sup> This recommendation appears for the first time on the inside of the back cover of *Psychological Science*, 16(12), December 2005. It has remained there unchanged ever since.

<sup>4</sup> Killeen (2005c) uses the symbol  $N$  to denote the cumulative distribution function of a standard normally distributed random variable. We use the Greek letter  $\Phi$ .

It seems that both Killeen and Cumming were alert to the potential ambiguity in how to compute the value of  $p_{\text{rep}}$  from a given  $p$ -value, but their recommendations were buried in an Appendix (Killeen, 2005a) and a Table caption (Cumming, 2005).

In any event a little thought shows that the correct connection between  $p_{\text{rep}}$  and the  $p$ -value from a one-sided test is not Eq. (2) but rather

$$p_{\text{rep}} = \Phi \left[ \frac{\Phi^{-1}(\max\{p, 1-p\})}{\sqrt{2}} \right]. \quad (5)$$

Calculation of  $p_{\text{rep}}$  must yield a number in the interval  $[\frac{1}{2}, 1]$  by its very definition as a posterior predictive probability (and recall Eq. (3) for explicit confirmation);  $p_{\text{rep}}$  never takes values below  $\frac{1}{2}$  and both Eqs. (4) and (5) respect this restriction. Allowing  $p_{\text{rep}}$  to take values in  $[0, \frac{1}{2})$ , as is permitted under Eq. (2), is to invite a jarring collision between what  $p_{\text{rep}}$  is intended to report and what it does in fact report.

Suppose prior to an experiment you have convinced yourself that the outcome will reflect a negative true effect parameter  $\delta$ , and you envisage a one-sided test of  $H_0 : \delta \geq 0$  vs.  $H_1 : \delta < 0$ . Your observed effect  $d$  turns out to be positive, contrary to expectations, and the one-sided  $p$ -value is 0.88. Eq. (2) gives  $p_{\text{rep}} = .20$ .

Now this can only mean the following: you have observed an experimental effect that disagrees with expectations. Despite the evidence, you are fairly sure ( $1 - p_{\text{rep}} = .80$ ) that a repetition of the experiment will yield a *negative* effect, in conflict with the data at hand but in agreement with your hypothesis. In other words, the evidence at hand has been overridden by your prior expectations and your view of the matter is supported by a *small* value of  $p_{\text{rep}}$ , and the smaller the better! Note that Eq. (5) gives the answer that is intended of a sensible posterior predictive probability of concurrence, namely  $p_{\text{rep}} = .80$ . The observed effect is positive and one has a legitimate Bayesian right to anticipate that a replication is more likely than not to produce a positive effect. We hasten to add, however, that this Bayesian prediction is by no means guaranteed to mirror the aleatory behavior of empirical replications. For a more detailed discussion of the critical distinction between substantive empirical replication and posterior predictive replication, consult the fourth and fifth Fits.

One might have expected that contributors to *Psychological Science*, not to mention reviewers and action editors, would have spotted the difficulty of interpretation that is built into Eq. (2) when a  $p$ -value exceeds  $\frac{1}{2}$ , and to have corrected the matter by reporting the complement. Perhaps some did so, but certainly others did not; even Sanabria and Killeen (2007) quote a value of  $p_{\text{rep}}$  below  $\frac{1}{2}$ . In Killeen (2005a, Figure 3) the trade-off between  $p_{\text{rep}}$  and the (one-sided)  $p$ -value based on Eq. (1) is abruptly cut off at  $p_{\text{rep}} = \frac{1}{2}$ , inviting the reader to interpret the tradeoff for  $p$ -values greater than  $\frac{1}{2}$ .

**FIT THE THIRD:** In which  $p_{\text{rep}}$  is misapplied

“Thats a great deal to make one word mean,” Alice said in a thoughtful tone. “When I make a word do a lot of work like that, said Humpty Dumpty, I always pay it extra.”

Through the Looking-Glass: Humpty Dumpty, Lewis Carroll, 1872.

The (incorrect) formula Eq. (2) for computing  $p_{\text{rep}}$  invites the unwary to carry out the indicated calculation whenever a  $p$ -value is available, regardless of the context in which the  $p$ -value arose. But it is wise to recall from the first Fit that  $p_{\text{rep}}$  is a posterior probability of *concurrence*, and that last term requires for its very meaning the notion of *sign* or direction of effect. What is the (unambiguous) direction associated with an interaction in a  $3 \times 4$  ANOVA, or for that matter the fact that the main effect of each variable is significant? More generally, what sense

of direction of effect is indicated by the fact that one cognitive model outperforms another on some body of data, as considered by Ashby and O'Brien (2008). As a careful reading of Ashby and O'Brien (2008) shows, their notion of replicability amounts to conventional power or something very similar. Many authors (e.g., Greenwald, Gonzalez, Guthrie, and Harris (1996), Oakes (1986) and Tversky and Kahneman (1971)) earlier used power as a means of quantifying ‘replicability’. But power, the complement of a Neyman–Pearson long-term error rate, is antithetical to Killeen’s views on statistical inference: “but once  $p_{\text{rep}}$  is determined, calculation of traditional significance is a step backward” (Killeen, 2005a, p. 349).

While we are on the topic of power, it is noteworthy that  $p_{\text{rep}}$  can be viewed as a predictive power calculation. One natural interpretation of predictive power is given in the following definition and calculation<sup>7</sup>:

$$\beta(\alpha, d) = \Pr \left( d^{\text{rep}} \sqrt{\frac{n}{2}} \operatorname{sgn} d \geq z_{\alpha} \mid d \right) = \Phi \left( \frac{|d| \sqrt{n/2} - z_{\alpha}}{\sqrt{2}} \right)$$

and it is seen at once that for  $\alpha = \frac{1}{2}$ ,  $\beta(\frac{1}{2}, d) = p_{\text{rep}}$ . In plain words, when significance means concurrence (and this is achieved when  $\alpha = \frac{1}{2}$ ),  $p_{\text{rep}}$  is predictive power. The trade-off between Type I and Type II errors ensures that a large value of  $\alpha$  is accompanied by a boost in power. No wonder then that  $p_{\text{rep}}$  so often returns large values that can mislead the casual consumer (see the fourth and fifth Fits for further detailed discussion).

As a concrete numerical example, suppose you have obtained an experimental effect  $d = 0.56$  based on a sample size  $n = 25$ . One computes predictive power = .59 and this provides but modest confidence that a replication will be significant at  $\alpha = .05$  (in the same direction as the original). In contrast  $p_{\text{rep}} = .92$ . The message conveyed by predictive power seems somewhat cautious in the first case ( $\alpha = .05$ ) but quite optimistic in the second ( $\alpha = .5$ ). The inflated confidence expressed by  $p_{\text{rep}}$  is revealed as a *legerdemain* arising from the mere shift of a decimal point.

Recently *Psychological Science* seems to have realized that the calculation of  $p_{\text{rep}}$  must be confined to its original scope, the simple two independent groups design, and that it does not readily extend beyond that limited scope (it does however extend to linear contrasts in ANOVA and to some analogous problems in regression). It is becoming increasingly common for the same author to report  $p_{\text{rep}}$  in a two-group comparison, but to switch to  $p$ -value for all other tests.<sup>8</sup> This is terribly awkward, and anyway prompts the question: why not report  $p$ -values for *all* tests, as was done routinely before the  $p_{\text{rep}}$  era? The answer of course is that, for a variety of good reasons,  $p$ -values themselves have been regarded as unsatisfactory and misleading. Wagenmakers (2007) gives an extensive review of the many shortcomings of  $p$ -values that have been exposed and discussed at length in the literature.

We thus discover that  $p_{\text{rep}}$  is not only equally unsatisfactory as the  $p$ -value when used as a test statistic, it is at the same time considerably more restricted in its scope and interpretation as an object of evidentiary import.

**FIT THE FOURTH:** In which  $p_{\text{rep}}$  invites mistrust

“I quite agree with you”, said the Duchess; and the moral of that is – ‘Be what you would seem to be’ – or, if you’d like to put it more simply—‘Never imagine yourself not to be otherwise than what it might appear to others that you were or might have been was not otherwise than what you had been would have appeared to them to be otherwise’.

<sup>7</sup> The *signum* function, abbreviated *sgn*, indicates the sign (+1 or –1) of a real variable. It is convenient to adopt the convention that  $\operatorname{sgn}(0) = 1$ .

<sup>8</sup> In his final editorial (Cutting, 2007) mixes  $p_{\text{rep}}$  and  $p$ -value without comment.

Alice's Adventures in Wonderland: The Mock Turtle's Story. Lewis Carroll, 1865.

Killeen (2005a, p. 349) discusses  $p_{\text{rep}}$  as a statistical estimator, saying

“As is the case for all statistics, there is sampling variability associated with  $p_{\text{rep}}$ , so that any particular value of  $p_{\text{rep}}$  may be more or less representative of the values found by other studies executed under similar conditions. *It is an estimate*”. [emphasis added].

The leading question is: What exactly is  $p_{\text{rep}}$  estimating? Addressing this question brings out the large variability of  $p_{\text{rep}}$  that all too frequently produces large numerical values, giving a naive consumer a misleading and exaggerated sense of optimism that a repetition of an experiment will concur with a given one.

Suppose you know the value of the population effect parameter  $\delta$ . You have in hand an observed effect  $d$  based on a per-group sample size  $n$ . Suppose a repetition of your experiment yields an independent observed effect  $d^{\text{rep}}$ . What is the probability that the two effects agree in sign (concur)? An elementary calculation gives

$$\begin{aligned} \Pr(d^{\text{rep}} \text{ concurs with } d \mid d, \delta) &= \Pr(dd^{\text{rep}} \geq 0 \mid d, \delta) \\ &= \Phi\left(\delta \operatorname{sgn} d \sqrt{\frac{n}{2}}\right) \\ &= \Phi(\Delta \operatorname{sgn} d). \end{aligned} \tag{6}$$

Here and below it is often convenient to write  $\Delta = \delta\sqrt{n/2}$ ;  $\Delta$  is a ‘non-centrality’ parameter, which determines power, familiar from classical inference. Note that if  $d$  and  $\delta$  disagree in sign, you would base your prediction on the sign of  $\delta$ , *not* on the sign of  $d$ , and your predictive probability (Eq. (6)) would be less than  $\frac{1}{2}$ . This stands in contrast to the prediction afforded by  $p_{\text{rep}}$  that relies on the sign of  $d$ , and which takes on values that are necessarily larger than  $\frac{1}{2}$ . We often abbreviate  $\Pr(d^{\text{rep}} \text{ concurs with } d \mid d, \delta)$  as  $\Pr(\text{concur} \mid d, \delta)$ .

Of course one does not know  $\delta$ , and it seems natural therefore to *estimate*  $\Pr(\text{concur} \mid d, \delta)$ .  $p_{\text{rep}}$  is the estimator proposed by Killeen (2005a) to do the job. We note that  $\Pr(\text{concur} \mid d, \delta)$  can be viewed – though very differently – from both a Bayesian and a frequentist standpoint and we discuss each interpretation in turn.

For Bayesians, it is natural to consider  $\Pr(\text{concur} \mid d, \delta)$  as a function of posterior belief  $f(\delta \mid d)$ . Indeed we have, from Killeen (2005a,b,c), Sanabria and Killeen (2007); and especially Cumming (2005), Doros and Geier (2005), and Macdonald (2005),

$$p_{\text{rep}} = E[\Pr(\text{concur} \mid d, \delta)] = \int \Pr(\text{concur} \mid d, \delta) f(\delta \mid d) d\delta, \tag{7}$$

in which the expectation is taken over the posterior distribution<sup>9</sup> of  $\delta$ . On the other hand it is unlikely that Bayesians would routinely summarize their posterior belief concerning  $\Pr(\text{concur} \mid d, \delta)$  by computing a single number such as  $p_{\text{rep}}$  or alternatively  $1 - p/2$  (which, as it happens, is the *median* value of  $\Pr(\text{concur} \mid d, \delta)$ ), when the entire posterior distribution of belief is available. If a summary measure is desired it is more informative to give a credible interval of values. In particular, the inequalities

$$\Phi\left(|d| \sqrt{\frac{n}{2}} - z_\alpha\right) \leq \Pr(\text{concur} \mid d, \delta) \leq 1,$$

give the endpoints for the  $(1 - \alpha)$  100% highest probability density (HPD) credible interval. For example,  $d = .56$  and  $n = 25$  yields

<sup>9</sup> If one adopts a flat prior on  $\delta$  (i.e.,  $f(\delta) \propto 1$ ), it is well known that the posterior density of  $\delta \mid d$  turns out to be normal with mean  $d$  and variance  $2/n$ . The integral in Eq. (7) is then straightforward and gives the standard expression in Eq. (3) for  $p_{\text{rep}}$ .

**Fig. 3.** An example of the density of the posterior random variable  $\Phi(\Delta \operatorname{sgn} d \mid d)$ , calculated using Eq. (8). Also shown are  $p_{\text{rep}}$ , which is the mean of the distribution,  $1 - p/2$ , which is the median, and  $p/2$ , which is the area under the density from 0 to 0.5.

$p_{\text{rep}} = .92$  and  $1 - p/2 = .976$ . In contrast, the 95% HPD credible interval is the rather modest prediction  $.63 \leq \Pr(\text{concur} \mid d, \delta) \leq 1$ . This broad credible interval for  $\Pr(\text{concur} \mid d, \delta)$  comes about because, regarded as a function of the random variable  $\delta \mid d$ , the probability density of  $\Pr(\text{concur} \mid d, \delta) = \Phi(\delta\sqrt{n/2} \operatorname{sgn} d)$  is strongly skewed towards  $\frac{1}{2}$ , as shown in Fig. 3. In other words, there is considerable posterior uncertainty about the probability that a future effect will concur with an original. A very similar and equally undesirable skew attends the predictive density of  $p$ -values (Cumming, *in press*), and essentially for the same reasons.

An example of the density of  $\Phi(\delta\sqrt{n/2} \operatorname{sgn} d \mid d)$  is shown in Fig. 3. The analytic form is as follows: for  $0 \leq t \leq 1$

$$f(t) = \exp(\Phi^{-1}(t) |z|) \exp(-z^2/2), \tag{8}$$

in which  $z = d\sqrt{n/2}$  is the  $z$ -score corresponding to the observed effect  $d$ . This density first appeared as a histogram based on a small-scale simulation in Cumming (2005). The most striking feature of the density is the large negative skew that is responsible for broad credible intervals

Another figure helps to explain why  $p_{\text{rep}}$  is often quite large, (e.g.  $p_{\text{rep}} \geq .85$ ), even though the true state of nature  $\delta$  is quite small and is thus likely to generate many more effects that conflict with an original than are predicted by  $1 - p_{\text{rep}}$ . In Fig. 4 an observed value of  $d$  is imagined to arise from a value of  $\delta$  that with probability  $\frac{1}{2}$  is larger than  $d$ , and with probability  $\frac{1}{2}$  is smaller. Three replications that might arise under a value of  $\delta$  that exemplifies each possibility are shown as open circles.  $p_{\text{rep}}$  is computed as a weighted average over all such imagined scenarios, the weights being provided by the posterior distribution  $f(\delta \mid d)$ .

Fig. 4 makes it clear that averaging over posterior uncertainty in  $\delta$  will often produce large values for  $p_{\text{rep}}$ , mainly because  $\Pr(\text{concur} \mid d, \delta) \approx 1$  when  $\delta > d$ , even though the true state of nature might be more like the one shown in the lower branch for which replicates can frequently be negative, in conflict with the original.

For a frequentist  $\delta$  is unknown but fixed, and as a statistic (i.e., as a function on the sample space)  $\Pr(\text{concur} \mid d, \delta) = \Phi(\Delta \operatorname{sgn} d)$  is the following dichotomous random variable:

$$\Pr(\text{concur} \mid d, \delta) = \begin{cases} \Phi(\Delta) & \text{with probability } \Phi(\Delta) \\ \Phi(-\Delta) & \text{with probability } \Phi(-\Delta). \end{cases} \tag{9}$$

The value of  $\Phi(|\Delta|)$  of  $\Pr(\text{concur} \mid d, \delta)$  arises whenever  $d$  and  $\delta$  concur; the value  $\Phi(-|\Delta|) = 1 - \Phi(|\Delta|)$  arises if  $d$  and  $\delta$  conflict in sign. Note that of the two values  $\Phi(\Delta)$  and  $\Phi(-\Delta)$  one is necessarily  $\geq \frac{1}{2}$  whereas the other is  $\leq \frac{1}{2}$ .





