Journal of Mathematical Psychology 53 (2009) 195–202

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

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

Geoffrey J. Iverson [a,b,*], Michael D. Lee [a,b], Shunan Zhang [a,b], Eric-Jan Wagenmakers [c]

[a] Department of Cognitive Sciences, University of California, Irvine, United States

[b] Institute for Mathematical and Behavioral Sciences, University of California, Irvine, United States

[c] Department of Psychology, University of Amsterdam, Netherlands

ARTICLE INFO

*Article history:*
Received 3 June 2008
Available online 22 November 2008

*Keywords:*
$p_{rep}$
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 $p_{rep}$ in lieu of the traditional $p$-value. In a polemic comprising five Fits we argue that $p_{rep}$ 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 $p_{rep}$ to the psychological community. He describes $p_{rep}$ as follows:

> "The statistic $p_{rep}$ 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, $p_{rep}$ 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 $p_{rep}$ when reporting their statistical analyses.

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

For its calculation, $p_{rep}$ 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.[2] All measurements are assumed to be mutually independent and normally distributed,

* 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 (G.J. Iverson), mdlee@uci.edu (M.D. Lee), szhang@uci.edu (S. Zhang), ej.wagenmakers@gmail.com (E.-J. Wagenmakers).

0022-2496/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.jmp.2008.09.004

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

[2] 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.

with a common known[3] 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 $d \sim N\left(\delta, \frac{2}{n}\right)$ 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

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

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.

[3] 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.

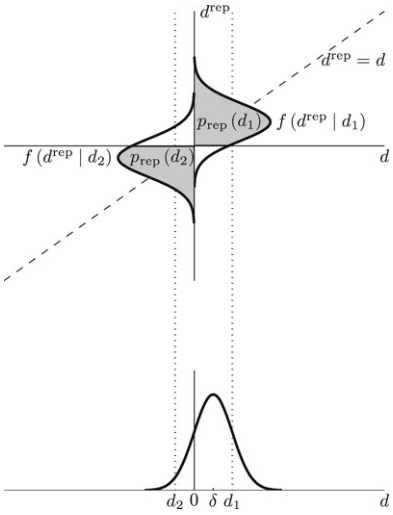


**Fig. 1.** Two independent experimental effects $d_1$ and $d_2$ are drawn from the distribution $f(d \mid \delta)$ generating the data. Each draw gives rise to a different value of $p_{\text{rep}}$, shown by shaded areas. Note that if the true state of nature $\delta$ is close enough to zero, $d_1$ 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 $M_1$:

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

The calculation of $p_{\text{rep}}$ is pictured in Fig. 1. It is the larger of the areas subtended by the posterior predictive $f(d^{\text{rep}} \mid d)$ above and below zero. Since $f(d^{\text{rep}} \mid d)$ 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

**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)[4]:

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

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

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

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

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

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[6]:

> "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

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

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

[5] An explicit calculation is indicated below in Eq. (7).

[6] 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.

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

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

Calculation of $p_{\text{rep}}$ must yield a number in the interval $\left[\frac{1}{2}, 1\right]$ 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 $\left[0, \frac{1}{2}\right)$, 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[7]:

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

and it is seen at once that for $\alpha = \frac{1}{2}, \beta\left(\frac{1}{2}, d\right) = p_{\text{rep}}$. 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.[8] 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".

[7] The *signum* function, abbreviated sgn, indicates the sign ($+1$ or $-1$) of a real variable. It is convenient to adopt the convention that $\text{sgn}(0) = 1$.

[8] 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

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

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 $\Pr\left(d^{\text{rep}} \text{ concurs with } d \mid d, \delta\right)$ as $\Pr\left(\text{concur} \mid d, \delta\right)$.

Of course one does not know $\delta$, and it seems natural therefore to *estimate* $\Pr\left(\text{concur} \mid d, \delta\right)$. $p_{\text{rep}}$ is the estimator proposed by Killeen (2005a) to do the job. We note that $\Pr\left(\text{concur} \mid d, \delta\right)$ 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 $\Pr\left(\text{concur} \mid d, \delta\right)$ as a function of posterior belief $f\left(\delta \mid d\right)$. Indeed we have, from Killeen (2005a,b,c), Sanabria and Killeen (2007); and especially Cumming (2005), Doros and Geier (2005), and Macdonald (2005),

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

in which the expectation is taken over the posterior distribution[9] of $\delta$. On the other hand it is unlikely that Bayesians would routinely summarize their posterior belief concerning $\Pr\left(\text{concur} \mid d, \delta\right)$ by computing a single number such as $p_{\text{rep}}$ or alternatively $1 - p/2$ (which, as it happens, is the *median* value of $\Pr\left(\text{concur} \mid d, \delta\right)$), 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

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

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

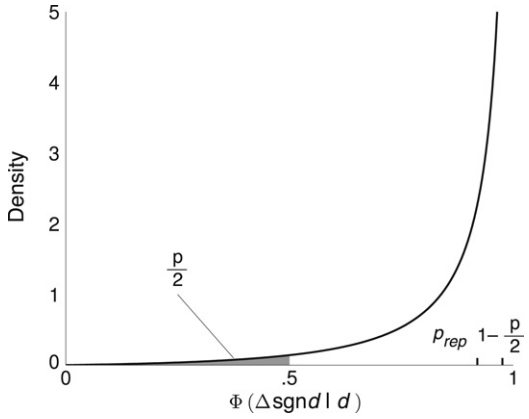


**Fig. 3.** An example of the density of the posterior random variable $\Phi\left(\Delta \text{ sgn } d \mid d\right)$, calculated using Eq. (8). Also shown are $p_{\text{rep}}$, which is the mean of the distribution, $1 - p/2$, which is the median, and $p/2$, which is the area under the density from 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 $.63 \leq \Pr\left(\text{concur} \mid d, \delta\right) \leq 1$. This broad credible interval for $\Pr\left(\text{concur} \mid d, \delta\right)$ comes about because, regarded as a function of the random variable $\delta \mid d$, the probability density of $\Pr\left(\text{concur} \mid d, \delta\right) = \Phi(\delta\sqrt{n/2}\,\text{sgn}\, d)$ 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 $\Phi(\delta\sqrt{n/2}\,\text{sgn}\, d \mid d)$ is shown in Fig. 3. The analytic form is as follows: for $0 \leq t \leq 1$

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

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 $f\left(\delta \mid d\right)$.

Fig. 4 makes it clear that averaging over posterior uncertainty in $\delta$ will often produce large values for $p_{\text{rep}}$, mainly because $\Pr\left(\text{concur} \mid d, \delta\right) \approx 1$ 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) $\Pr\left(\text{concur} \mid d, \delta\right) = \Phi\left(\Delta \text{ sgn } d\right)$ is the following dichotomous random variable:

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

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

[9] If one adopts a flat prior on $\delta$ (i.e., $f\left(\delta\right) \propto 1$), 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 $p_{\text{rep}}$.

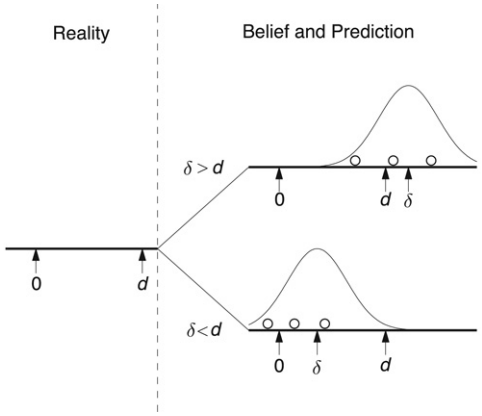


**Fig. 4.** Given an observed effect $d$, the possibility that $\delta > d$ is exemplified in the upper scenario on the right, which shows three independent replicate effects as open circles. Equally likely is the possibility that $\delta < d$ and a typical scenario is depicted in the lower branch. $p_{\text{rep}}$ is computed as a weighted average over all such scenarios, the weights being provided by the posterior distribution $f(\delta \mid d)$.

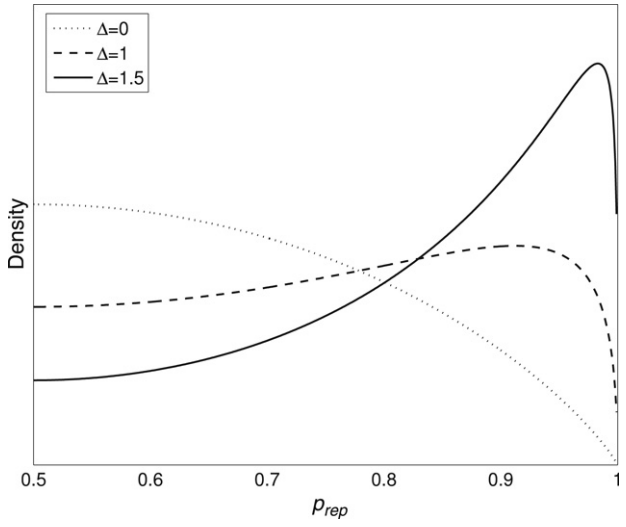


**Fig. 5.** The sampling density of $p_{\text{rep}}$ for $|\Delta| = 0$, 1, and 1.5.

$p_{\text{rep}} = \Phi(|d|\sqrt{n/4})$ is concentrated on $\left[\frac{1}{2}, 1\right]$, and as an estimator for $\Pr(\text{concur} \mid d, \delta)$, which takes values in $[0, 1]$, its inability to take values in $\left[0, \frac{1}{2}\right)$ presents it with a very difficult challenge. This restriction on range is especially worrisome for small-to-moderate values of $|\Delta|$. For example if $|\Delta| \approx 0$ the values $\Phi(\Delta)$ and $\Phi(-\Delta)$ are each approximately $\frac{1}{2}$ whereas we know from its null distribution that the median value of $p_{\text{rep}}$ is $\Phi\left(\sqrt{1/2}\Phi^{-1}(3/4)\right) \approx .68$ and its expected value is $\frac{1}{2} + \frac{1}{\pi}\arcsin(\sqrt{1/3}) \approx .70$.

In Fig. 5 we plot the distribution of $p_{\text{rep}}$ for several values of $\Delta$. These functions are members of the following family of expressions indexed by $|\Delta|$:

$$f_{p_{\text{rep}}}(t) = 2\sqrt{2}\exp\left[-\Delta^2/2\right]\cosh\left[\sqrt{2}\Phi^{-1}(t)\,\Delta\right] \times \exp\left[-\left(\Phi^{-1}(t)\right)^2/2\right], \quad \frac{1}{2} \le t \le 1.$$

The null distribution corresponds to $\Delta = 0$.

For small values of $|\Delta|$, $p_{\text{rep}}$ overestimates its target, often by a large amount. On the other hand, for a sufficiently large value of $|\Delta|$ one has $\Phi(|\Delta|) \approx 1$, whence $p_{\text{rep}} \approx 1$ as well. This last rather trite fact does not, however, make $p_{\text{rep}}$ a particularly good estimator for $\Pr(\text{concur} \mid d, \delta)$ even though both probabilities approximate 1. As we will see in Fig. 6, for large $|\Delta|$, $p_{\text{rep}}$ systematically underestimates $\Pr(\text{concur} \mid d, \delta)$ and the silly estimator 1 will often do a better job.

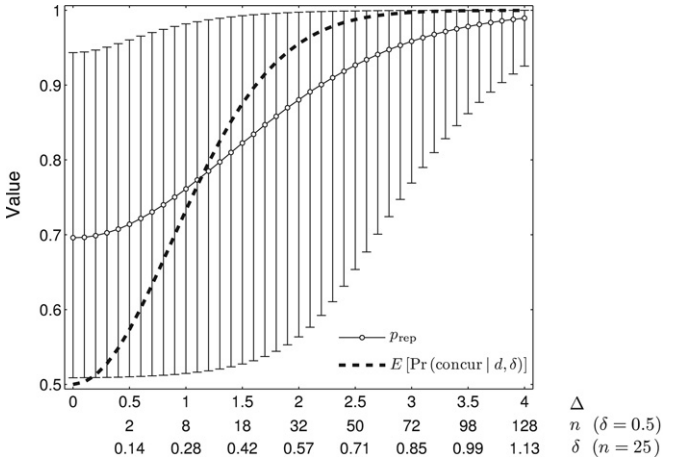


**Fig. 6.** The expected values of $p_{\text{rep}}$ and $\Pr(\text{concur} \mid d, \delta)$ as functions of the non-centrality parameter $\Delta$, with equivalent representative values for effect size $\delta$ and sample size $n$ shown. The systematic bias in $p_{\text{rep}}$ accounts for the difference between the two curves. The large variability in $p_{\text{rep}}$ is evident in the large error bars that represent the 95% equal area intervals.

One can understand visually the poor performance of $p_{\text{rep}}$ by plotting its expected value and the long-run expected value of $\Pr(\text{concur} \mid d, \delta)$, which equals $\Phi^2(|\Delta|) + \Phi^2(-|\Delta|)$, on the same axes, as functions of $\Delta$. These plots are given in Fig. 6. It is evident that the expected value $E\left[p_{\text{rep}}\right]$ is much larger than $E[\Pr(\text{concur} \mid d, \delta)]$ for small values of $\Delta$, but is dominated by it for larger values of $\Delta$. The error bars represent the 95% equal area intervals of the sampling distribution of $p_{\text{rep}}$. The bias and imprecision of $p_{\text{rep}}$ as an estimator is evident for all but large effects or sample sizes (values of $\Delta$ in excess of 3.5), for which it close to 1.

**FIT THE FIFTH**: The *Psychological Science* action editor's dilemma

> He was thoughtful and grave—but the orders he gave
> Were enough to bewilder a crew.
> When he cried "Steer to starboard, but keep her head larboard!"
> What on earth was the helmsman to do?
>
> The Hunting of the Snark: FIT THE SECOND, The Bellman's Speech. Lewis Carroll, 1876.

A fundamental distributional difference underlies the construction of $p_{\text{rep}}$ and $\Pr(\text{concur} \mid d, \delta)$. Under model $M_1$ substantive effects are independent draws from a normally distributed population, with mean $\delta$ and variance $2/n$. These effects are what action editors examine when replications of an original experiment come across their desks; these draws, provided by science, determine $\Pr(\text{concur} \mid d, \delta)$. On the other hand, $p_{\text{rep}}$ is the probability of an event involving values of $d^{\text{rep}} \mid d$ and those values are draws from a normally distributed predictive distribution, with mean $d$ and variance $4/n$. Values of $d^{\text{rep}} \mid d$ are not substantive replicates, and they are not independent. To confuse them with independent, substantive replicates is a mistake.

With this remark in mind we now examine how $p_{\text{rep}}$ can misinform the important business of scientific induction that is carried out daily by authors, reviewers and action editors.

Suppose you are an action editor for *Psychological Science* and a paper for review comes across your desk that reports a surprising and somewhat controversial finding. The evidence for the effect in question is summarized in the following data: $d = .56$, $n = 25$, $z = 1.98$, $p(\text{two-sided}) = .05$, $p_{\text{rep}} = .92$. If this finding is true it will wrinkle the theoretical cloth of an important branch

of experimental psychology, and likely promote new directions for empirical and theoretical research.

The referees are enthusiastic and you accept the article for publication. A few months after publication, independent experimental replications make an appearance. In fact the first three lie unopened on your desk. Before opening any one, you ask yourself what you expect on the basis of $p_{\text{rep}}$. All three replications might exhibit positive effects, or none might. You quickly tabulate the various possibilities and the accompanying probabilities as determined by $p_{\text{rep}}$ and the Binomial distribution.[10] You find the probabilities for 0, 1, 2 and 3 concurrences to be .00, .02, .20 and .78, respectively.

Armed with your calculations, you are confident that at least two of the three replications will concur with the original, and you would not be surprised if all three did so. Opening the new submissions you are dismayed to discover that two of the three articles report negative effects, in conflict with the original. The relevant data are: $d_1^{\text{rep}} = 0.40, n = 25, z = 1.41, p$ (two-sided) $= .16, p_{\text{rep}} = .84; d_2^{\text{rep}} = -0.08, n = 25, z = -0.28, p$ (two-sided) $= .78$, $p_{\text{rep}} = .58$; and $d_3^{\text{rep}} = -0.03$, $n = 25$, $z = -0.11$, $p$ (two-sided) $= .91, p_{\text{rep}} = .53$.

The authors who obtained the positive effect $d^{\text{rep}} = 0.40$ claim a replication of the original, despite the rather large $p$-value, and in their discussion call for an aggressive experimental foray along lines suggested by the original finding. In marked distinction, the authors who found small negative effects are quite critical of the original finding and state quite clearly that their efforts to replicate had failed utterly, and that little purpose would be served by pursuing this particular line of research.

As action editor how are you to react to these unexpected and conflicting findings? Did something quite unusual occur, or is there a subtle causal artifact at work that would explain the two negative outcomes, one that you suspect will be hard if not impossible to uncover.

Neither of these reactions is warranted. In fact the data are quite consistent with one another and with the model $M_1$ that underlies the standard calculation of $p_{\text{rep}}$; and there are multiple considerations that support this view of the data. Based on the original data, the 95% posterior predictive interval for future replications is $[-0.22, 1.34]$ and this interval[11] readily accommodates all of the observed replicate effects, though it is not obliged to do so. A $\chi^2$ test of the assumption that all four effects are replicates (i.e., that all four are generated by a common value of $\delta$ yields $\chi^2\,(3) = 3.75$, and this $\chi^2$ value is not close to signaling any significant differences among the various experimental outcomes. Further, all data are compatible with a true value of $\delta$ about 0.20 (note that the arithmetic average of all four experimental effects is 0.21).[12] Assuming for illustration that $\delta = 0.20$, the probability that any single replicate effect will be negative is .23, and consequently the observed pattern of replications (2 negative, 1 positive) is expected to occur about 12% of the time; this last probability is 6 times larger than the corresponding binomial prediction based on $p_{\text{rep}}$. We remind the reader of the lower branch of Fig. 4.

## Postlude to the agony: Caveat emptor

> He had bought a large map representing the sea,
> Without the least vestige of land:
> And the crew were much pleased when they found it to be
> A map they could all understand.
>
> The Hunting of the Snark: FIT THE SECOND, The Bellman's Speech. Lewis Carroll, 1876

If you must use $p_{\text{rep}}$ do so with caution, a deliberate purpose in mind, and with full awareness of its shortcomings as an estimator of Pr (concur $\mid d, \delta$). As we have seen, $p_{\text{rep}}$ does not quantify 'replicability' of experimental effects, it does not appear to generalize beyond linear contrasts, and as an estimator of *concurrence* it is unreliable and is in fact not even consistent.[13]

To buy into $p_{\text{rep}}$ as it is currently promoted by *Psychological Science* is to buy into the *significance fallacy*, the belief that significant effects are highly reliable and replicable (Oakes, 1986; Tversky & Kahneman, 1971).[14] Not only does $p_{\text{rep}}$ encourage that erroneous belief, it sanctions it with the authority and precision of a quantitative calculation.

In particular do not use $p_{\text{rep}}$ as Psychological Science currently does, merely as a convenient way to lower the bar on conventional criteria for significance, allowing Type I errors to triple in frequency over conventional 5% rates, not to mention sanctioning a fourteen-fold increase over the more conservative (but often preferable) 1% standard. Despite the encouraging words that have been bandied about in praise of $p_{\text{rep}}$, the fact remains that $p_{\text{rep}} = .85$ corresponds to a $p$-value of .14 and $p_{\text{rep}} = .90$ corresponds to a $p$-value of .07. If, based on $p$-value of .14, you would not reject the possibility that $|\delta|$ is quite small, perhaps negligible, why would you offer much better than even odds that a substantive replication would agree in sign with an original?

You protest: surely 85% and 90% are much closer to 100% than they are to 50%. Our response is that this simple fact of arithmetic is as misleading as it is true (for reasons detailed in our fourth and fifth Fits). The probability scale provided by $p_{\text{rep}}$ (or any other probability value for that matter) is the wrong metric on which to evaluate evidence about $\delta$. If you had asked different questions of your data, for instance what does your value of $p_{\text{rep}}$ tell you about $\delta = 0$ versus $\delta \neq 0$, we would encourage you to compute a ratio of probabilities. The resulting Bayes Factor, a ratio of probability densities, is a sensible and readily interpretable means of evaluating (relative) evidence (e.g., Bernardo and Smith (1994), Kass and Raftery (1995) and Lee and Wagenmakers (2005)).[15] It should be the routine business of authors contributing to *Psychological Science* or any other Journal of scientific psychology to report Bayes Factors. Presently only a small handful do so, in stark contrast to current practice in the statistical community.

[10] Actually, this is not the way Bayesian posterior predictive probabilities are computed. However the correct calculations do not change the conclusions of this analysis.

[11] Both frequentists and Bayesians (assuming a flat prior on $\delta$) agree on the form of this predictive interval, though they differ considerably on its interpretation. The general form of the interval is $d - z_{\alpha/2}\sqrt{4/n} \leq d^{\text{rep}} \leq d + z_{\alpha/2}\sqrt{4/n}$.

[12] The symmetric 95% HPD credible interval for $\delta$ based on $d$ is [.01, 1.11]. Based on all four experimental effects it is $[-.067, 0.487]$.

[13] The term 'consistent' is standard in statistics (Casella & Berger, 2002). A sequence of estimators $\hat{\vartheta}_n$ is consistent for a parameter $\vartheta$ if, for every $\epsilon > 0$ and every $\vartheta$, $\lim_{n\to\infty} \Pr\left(\left|\hat{\vartheta}_n - \vartheta\right| \geq \epsilon\right) = 0$. This property, which is usually the very least one requires of an estimator, obviously does not hold for $p_{\text{rep}}$. When $\delta = 0$ we have Pr (concur $\mid d, \delta$) $= \frac{1}{2}$; and since, with probability 1, $p_{\text{rep}}$ does not take on the value $\frac{1}{2}$ it can hardly be said to be consistent for its target.

[14] The term 'significance fallacy' is our terminology. Oakes called the common but unjustified belief in the replicability of significant effects the 'significance hypothesis', whereas Tversky and Kahneman discussed the matter in terms of a folk-theorem, a 'law of small numbers', in a particular example of their more general study of representativeness.

[15] Under model $M_1$ and a flat prior on $\delta$, a Bayes Factor for selecting between the hypotheses $H_0 : \delta = 0$ and $H_1 : \delta \neq 0$ is given by $B_{01} = \sqrt{n}\exp\left(-z^2/2\right)$. When $d = 0.56$ and $n = 25$, one has $z = 0.56 \times 5/\sqrt{2} = 1.98$ and $B_{01} = 0.7$. Yes, the data favor $H_1$ over $H_0$, but by a factor that is scarcely worth the mention. If *a priori* you believed Pr $(H_0) = .5$, *a posteriori* you believe Pr $(H_0 \mid d) = .41$; if *a priori* you believed Pr $(H_0) = .1$ the data have modified your belief so that Pr $(H_0 \mid d) = .065$. In either case, the data $d = 0.56$ and $n = 25$ are pretty much inconsequential. What price then the frequentist asterisk and the declaration "significant at level .05"?

## Acknowledgments

We thank many colleagues for their comments at one stage or another in the preparation of this work: Bill Batcheler, Barbara Dosher, Jean-Claude Falmagne, Yung-Fong Hsu, R. Duncan Luce, Larry Maloney, Roger Ratcliff, Jeffrey Rouder, Ching-Fan Sheu, George Sperling. We are also most grateful to Sheng Kung (Mike) Yi and Si Yi Deng for early help with relevant calculations.

## 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*(1), 16–27.
Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.
Casella, G., & Berger, R. (2002). *Statistical inference* (Second ed). Duxbury: Pacific Grove.
Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, *16*(12), 1002–1004.
Cumming, G. (in press). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. Perspectives on Psychological Science.
Cutting, J. E. (2005). Acknowledgment. *Psychological Science*, *16*(12), 1013.
Cutting, J. E. (2007). Rhythms of research. *Observer*, *20*(11), 11–14.
Doros, G., & Geier, A. B. (2005). Probability of replication revisited: Comment on an alternative to null–hypothesis significance tests. *Psychological Science*, *16*, 1005–1006.
Greenwald, A. G., Gonzalez, R., Guthrie, D. G., & Harris, R. J. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology*, *33*, 175–183.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 377–395.
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.
Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.
Macdonald, R. R. (2005). Why replication probabilities depend on prior probability distributions: A rejoinder to Killeen (2005). *Psychological Science*, *16*(12), 1007–1008.
Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Chichester: Wiley.
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.
Tversky, A., & Kahneman, D. (1971). The belief in the 'law of small numbers'. *Psychological Bulletin*, *76*, 105–110.
Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804.