A Model Averaging Approach to Replication: The Case of p_{rep}

Geoff Iverson[1], Eric–Jan Wagenmakers[2], and Michael D. Lee[1]
[1] University of California Irvine
[2] University of Amsterdam

Correspondence concerning this article should be addressed to:
Eric-Jan Wagenmakers
University of Amsterdam, Department of Psychology
Roetersstraat 15
1018 WB Amsterdam, The Netherlands
Ph: (+31) 20–525–6420
Fax: (+31) 20-639-0279
E-mail may be sent to EJ.Wagenmakers@gmail.com.

Abstract

The purpose of the recently proposed p_{rep} statistic is to estimate the probability of concurrence, that is, the probability that a replicate experiment yields an effect of the same sign (Killeen, 2005a). The influential journal *Psychological Science* endorses p_{rep} and recommends its use over that of traditional methods. Here we show that p_{rep} overestimates the probability of concurrence. This is because p_{rep} was derived under the assumption that all effect sizes in the population are equally likely *a priori*. In many situations, however, it is advisable to also entertain a null hypothesis of no or approximately no effect. We show how the posterior probability of the null hypothesis is sensitive to *a priori* considerations and to the evidence provided by the data; and the higher the posterior probability of the null hypothesis, the smaller the probability of concurrence. When the null hypothesis and the alternative hypothesis are equally likely *a priori*, p_{rep} may overestimate the probability of concurrence by 30% and more. We conclude that p_{rep} provides an upper bound on the probability of concurrence, a bound that brings with it the danger of having researchers believe that their experimental effects are much more reliable than they actually are.

Keywords: Statistical Hypothesis Testing, Prediction, Model Averaging, Bayesian Analysis

Suppose you conduct an experiment to test whether words such as *pizza* prime words such as *coin* (Pecher, Zeelenberg, & Raaijmakers, 1998). The motivating hypothesis states that prior presentation of a word facilitates later processing for another word when both words refer to objects with similar physical attributes (e.g., pizzas and coins are both round and flat). This relatively little studied "perceptual priming" effect may be contrasted with the well–established "associative priming" effect in which the presentation of a word such as *butter* facilitates later processing for the associatively related word *bread* (Meyer & Schvaneveldt, 1971).

In your priming experiment, you measure the effects of perceptual priming and associative priming, and a p–value hypothesis test shows that both effects are significant—by coincidence, both tests yield $p = .03$. Blissfully unaware of the work by Pecher et al. (1998), you set out to submit a report to one of psychology's premier journals, *Psychological Science*. You browse the journal's Author Guidelines and find that authors are encouraged to replace the traditional p value with Peter Killeen's p_{rep} value (Ashby & O'Brien, 2008; Killeen, 2005, 2005a, 2005b, 2006, 2007; Sanabria & Killeen, 2007; for discussion see Cumming, 2005; Doros & Geier, 2005; Macdonald, 2005; Wagenmakers & Grünwald, 2006).

Because you do not want to decrease needlessly your chances of getting accepted by *Psychological Science*, you transform your two–sided p–values of $p = .03$ into p_{rep} values of approximately .94 (e.g., Killeen, 2005, p. 17; Killeen, 2005a, p. 353). You do not have the time to analyze the statistics in the Killeen articles carefully, but you understand that you can conclude that, should you repeat the experiment, there is a $p_{rep} \approx .94$ probability of again finding each priming effect, even though the replication may not be statistically significant (i.e., replication refers to *concurrence*, that is, finding a replicate effect of the same sign; Killeen, 2005a, p. 346). Is this conclusion justified? We believe it is not.

Our disbelief stems from the fact that p_{rep} is based on a single model, namely the model that assumes all effect sizes to be equally likely *a priori* (Doros & Geier, 2005; Killeen, 2005b). We call this model H_1. The p_{rep} statistic does not take into account the plausibility of the simpler model, H_0, that postulates that an effect is either completely absent or so small that it would take a much larger sample for it to be detected. When H_0 is deemed plausible—either through *a priori* considerations or through the information provided by the data—this should considerably reduce one's confidence of finding concurrence, as $p_{rep} \approx 1/2$ under H_0 (see also Macdonald, 2005).[1]

To illustrate the impact of *a priori* considerations on the probability of concurrence, let us revisit the priming experiment and its $p_{rep} \approx .94$ for the established phenomenon of associative priming and the new phenomenon of perceptual priming. Imagine that you are given \$100 to bet that a replicate effect will concur with your data; that is, you get to keep the \$100 when the effect of your choice (i.e., either perceptual priming or associative priming) replicates, but you lose the \$100 that you were given when the effect of your choice does not replicate. In our priming example, the p_{rep} statistic suggests that you have no

[1] We write $p_{rep} \approx 1/2$ and not $p_{rep} = 1/2$ because our argument also holds when H_0 is only approximately true, as we later discuss in detail.

Correspondence concerning this article may be addressed to Eric–Jan Wagenmakers, University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands.

grounds to prefer one effect over the other, as $p_{rep} \approx .94$ for both. Nevertheless, the smart money in the betting scenario will be on the established effect, not on the new effect.

The reason for this is the very real possibility that your new finding of perceptual priming is a fluke, that is, a Type I error—and if this is the case, then p_{rep} is 1/2, not .94. On the other hand, associative priming has been reported in countless studies since Meyer and Schvaneveldt (1971), and this knowledge increases the probability that your measurement of this effect is real, which in turn increases the probability of finding concurrence in a replication of your priming experiment.

Thus, it is evident that established but not novel effects inspire high confidence of concurrence. This observation is, however, at odds with the standard interpretation of the p_{rep} statistic (see also Macdonald, 2005); in the following, we elaborate and formalize this intuitive argument and show how the calculation of the probability of concurrence requires one to take both H_0 and H_1 into account simultaneously.

The p_{rep} Statistic

Consider an experiment that features two conditions. Let d denote the observed effect size, and d_{rep} the effect size in a replication study. Concurrence is observed when d and d_{rep} have the same sign, so that $d_{rep}d \geq 0$. As explained in Doros and Geier (2005) and Killeen (2005a, 2005b), p_{rep} is computed under the assumption of a flat improper prior distribution[2] on the true population effect δ. *This assumption is H_1 in our terminology.* To be explicit,

$$p_{rep} = \Pr\left(d_{rep}d \geq 0 | d, H_1\right). \tag{1}$$

Note that this equation does not feature δ; that unobserved parameter has been integrated out using the flat improper prior distribution. Of course, one could argue whether such a flat distribution on δ is appropriate (Iverson, Lee, Zhang, & Wagenmakers, in press a; Iverson, Lee, & Wagenmakers, in press b): indeed, in Bayesian statistics, it is standard practice to use prior distributions that put more mass on small effect sizes than on large effect sizes.[3] For instance, the prior on effect size in the Zellner and Siow (1980) Bayesian hypothesis test is the Cauchy distribution (i.e., a t–distribution with one degree of freedom). Here we sidestep this discussion and wish to point out only that p_{rep} has Bayesian roots, and can therefore be given a Bayesian interpretation: p_{rep} estimates the probability that a replicate effect will concur with an original, under the assumption of a flat prior on the true population effect δ. More details are given in Appendix A.

The statistic p_{rep} can also be given a frequentist interpretation. Specifically, p_{rep} relates to the traditional two–sided p–value as (Killeen, 2005, p. 17):

$$p_{rep} = \Phi\left[\Phi^{-1}\left[1 - \frac{p}{2}\right]/\sqrt{2}\right], \tag{2}$$

where Φ denotes the standard Normal cumulative distribution function. Figure 1 plots the relation between p_{rep} and the two–sided p–value. As can be seen from the figure, the relation between the two statistics is close to linear, both for $p \in (0, 0.5)$ (left panel) and for $p \in (0, 0.1)$ (right panel). Divergence from linearity is only apparent when we consider p–values that are very small.

[2] An improper prior is a density function that does not integrate to a finite number.

[3] Despite the overlap in its conclusion and authors, the Iverson et al. (in press a) and Iverson et al. (in press b) articles differ from the current one in content and focus.

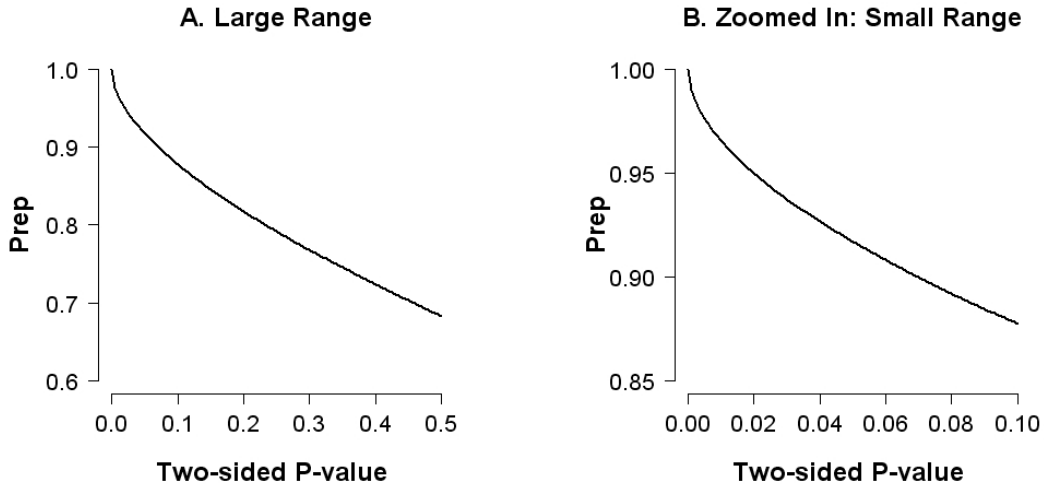

Figure 1. The function that relates p_{rep} to the two–sided p–value (Killeen, 2005, p. 17).

Given the almost linear mapping between the two–sided p–value and p_{rep}, one might well wonder to what extent p_{rep} provides information that the traditional p–value does not. Doros and Geier (2005, p. 1006) argued that "(...) any measure that is no more than a simple transformation of the classical p value (...) will inherit the shortcomings of that p value." In response, Killeen (2005b, p. 1011) argued that p_{rep} and the p–value, "although informationally equivalent, are distinguished by the inferences they warrant; p_{rep} is a valid posterior predictive probability, p is not."

Bayesian Model Averaging and its Effect on p_{rep}

As we have seen above, the statistic p_{rep} calculates the probability of concurrence under H_1, that is, under the assumption of a flat prior distribution for the population effect size δ. We propose to also consider an alternative model, H_0, that states that $\delta \approx 0$. Under H_0, the value of p_{rep} is $\Pr(d_{rep}d \geq 0|d, H_0) \approx 1/2$. We will later discuss the extent to which H_0 is plausible, but for now we proceed by noting that in standard statistical practice, be it Bayesian, frequentist, or likelihood–based, H_0 is generally considered plausible and may even be assigned special status.

Thus, we now have two estimates for the probability of concurrence, one under H_1 and one under H_0. How might we proceed? One way would be to settle on the estimate provided by the most likely model. However, this would mean that a small change in the data—say a minimal change that switches our preference from H_1 to H_0—could lead to a dramatic change in p_{rep}. More generally, a procedure that focuses on the estimates from a single model ignores the uncertainty in model selection, and therefore results in "(...) over–confident inferences and decisions that are more risky than one thinks they are" (Hoeting, Madigan, Raftery, & Volinsky, 1999, p. 382).

A second, better way to proceed is to construct a single estimate for the probability of concurrence, one that does not depend on the model that is being entertained. To this end,

one can calculate a weighted average of the two p_{rep}'s, in which the weights are given by the posterior probabilities of H_1 and H_0. This procedure is commonly known as *Bayesian model averaging* (e.g., Draper, 1995; Hoeting et al., 1999; Madigan & Raftery, 1994). Note that a small change in the data causes only a small change in the posterior model probabilities, so that, even though the preference order for the models may switch, the change in the model averaged estimate of p_{rep} will be small.

This means that when we apply the Bayesian model averaging procedure to p_{rep}, we should find that the weighted average of p_{rep} values yields a more realistic estimate of the probability of concurrence, an estimate that dampens the enthusiasm brought about by an analysis that only considers H_1. Specifically, the model averaged value for p_{rep}, denoted here by p_{rep}^{Bma} is always smaller or equal to that of the original p_{rep} which is conditional on H_1:

$$\begin{aligned} p_{rep}^{Bma} = \Pr\left(d_{rep}d \geq 0|d\right) &= \Pr\left(H_0|d\right)\Pr\left(d_{rep}d \geq 0|d, H_0\right) + \Pr\left(H_1|d\right)\Pr\left(d_{rep}d \geq 0|d, H_1\right) \\ &= \Pr\left(H_0|d\right) \times 1/2 + \Pr\left(H_1|d\right) \times p_{rep} \\ &\leq p_{rep}. \end{aligned} \tag{3}$$

The foregoing shows that, from a Bayesian perspective at least, it is prudent to calculate the overall probability of concurrence as a weighted average of the separate probabilities of concurrence under H_0 and H_1. Note that in order to arrive at an estimate for the weights—the posterior model probabilities—one needs to update the prior model probabilities $\Pr(H_0)$ and $\Pr(H_1)$ by means of the data $f(D|H_0)$ and $f(D|H_1)$ to posterior model probabilities $\Pr(H_0|D)$ and $\Pr(H_1|D)$, respectively. This implies that p_{rep}^{Bma} will be lower than p_{rep} to the extent that $\Pr(H_0|D)$ is large. This in turn occurs when prior considerations lead to a high value for the prior model probability $\Pr(H_0)$, or when the data are relatively likely under H_0, that is, when $f(D|H_0)$ is relatively large compared to $f(D|H_1)$.

The above analysis also clarifies why, in our earlier priming example, one would have more confidence in replication of the well-established effect than in that of the new effect; for the new effect, $\Pr(H_0)$ is relatively large, and this leads p_{rep}^{Bma} to be relatively small. To get a feeling for the extent to which model averaging drives down estimates for the probability of concurrence, we now turn to an example from a default Bayesian hypothesis test.

Illustration: Model Averaging for A Default Bayesian Hypothesis Test

Before presenting the results from the Bayesian hypothesis test, it is important to introduce some key concepts of Bayesian inference. More information can be found in Bayesian articles and books that discuss philosophical foundations (Lindley, 2000; O'Hagan & Forster, 2004), computational innovations (Gamerman & Lopes, 2006), and practical contributions (Congdon, 2003).

Assume you contemplate two models, H_0 and H_1, and seek to quantify model uncertainty in terms of probability. Consider first H_0. Bayes' rule dictates how your prior probability of H_0, $\Pr(H_0)$, is updated through the observed data D to give the posterior probability of H_0, $\Pr(H_0|D)$:

$$\Pr(H_0|D) = \frac{\Pr(H_0)f(D|H_0)}{\Pr(H_0)f(D|H_0) + \Pr(H_1)f(D|H_1)}. \tag{4}$$

In the same way, one can calculate the posterior probability of H_1, $\Pr(H_1|D)$. The ratio of these posterior probabilities is given by

$$\frac{\Pr(H_0|D)}{\Pr(H_1|D)} = \frac{\Pr(H_0)}{\Pr(H_1)} \frac{f(D|H_0)}{f(D|H_1)}. \tag{5}$$

This equation shows that the change from prior odds $\Pr(H_0)/\Pr(H_1)$ to posterior odds $\Pr(H_0|D)/\Pr(H_1|D)$ is determined entirely by the ratio of the marginal likelihoods $f(D|H_0)/f(D|H_1)$.[4] This ratio is generally known as the *Bayes factor* (Jeffreys, 1961), and the Bayes factor, or the log of it, is often interpreted as the *weight of evidence* coming from the data (Good, 1985). A hypothesis test based on the Bayes factor prefers the model under which the observed data are most likely (for details see Berger & Pericchi, 1996; Bernardo & Smith, 1994, Chapter 6; Gill, 2002, Chapter 7; Klugkist, Laudy, & Hoijtink, 2005; Kass & Raftery, 1995; O'Hagan, 1995). Note that the Bayes factor quantifies the evidence for H_0 versus H_1 without taking into account the prior plausibility of the models.

Having established the necessary terminology, we can now discuss the effect of model averaging on p_{rep} using a Bayesian Z–test that was proposed by Smith and Spiegelhalter (1980). This Bayesian Z–test is fully automatic, and, as far as automatic hypothesis tests go, its performance is regarded as "quite satisfactory" (Berger & Delampady, 1987, p. 319). The Smith and Spiegelhalter Z–test estimates the Bayes factor by means of the following equation:

$$B_{01} = \frac{f(D|H_0)}{f(D|H_1)} = \sqrt{n} \exp\left[-\frac{Z^2}{2}\right], \tag{6}$$

where $Z = d\sqrt{n/2}$ is the familiar frequentist test statistic.

From Equation 3 and the derivations in Appendix A, the model averaged probability of concurrence is given by

$$p_{rep}^{Bma} = \Pr(H_0|d) \times 1/2 + \Pr(H_1|d) \times \Phi\left[\frac{|Z|}{\sqrt{2}}\right], \tag{7}$$

where Φ again denotes the standard Normal cumulative distribution function. The posterior model probabilities in Equation 7 can be obtained from Equations 5 and 6. For example, suppose that an experiment with $n = 25$ yields $d = 0.56$. We compute $Z = 1.98$, and plugging this into Equation 6 yields $B_{01} \approx 0.704$. When H_0 and H_1 are equally likely a priori, $\Pr(H_0|d)$ is given by $B_{01}/(B_{01} + 1)$; in this case then, $\Pr(H_0|d) = 0.704/1.704 \approx .413$, and $\Pr(H_1|d)$ is its complement, $1 - .413 = .587$. The standard normal cumulative distribution function of $1.98/\sqrt{2}$ equals .919. Putting these separate components together in Equation 7 yields $p_{rep}^{Bma} = .413 \times 1/2 + .587 \times .919 = .746$. It is striking that .746, the model averaged probability of concurrence, is considerably more conservative than the .919 value that is obtained when H_0 is ignored (i.e., on a scale from .5 to 1, this amounts to a decrease of 34.6%).

To demonstrate more fully the effect of model averaging on the estimated probability of concurrence, Figure 2 shows three different p_{rep} methods as a function of Z–score. In each panel of Figure 2, the solid line gives Killeen's original p_{rep} estimate. The dashed and

[4] The likelihoods $f(D|H.)$ are called marginal because the model parameters have been integrated out.

dotted lines provide the model averaged estimates, that is, p_{rep}^{Bma}. The model weights (i.e., the posterior probabilities of H_1 and H_0) are based on the Smith and Spiegelhalter Bayesian Z–test. Other Bayesian tests, such as the one proposed by Zellner and Siow (1980), yield similar results. The dashed line corresponds to an H_1 that is *a priori* just as likely as H_0 (i.e., $\Pr(H_1) = \Pr(H_0) = 0.5$); the dotted line corresponds to an H_1 that is *a priori* less likely than H_0 (i.e., $\Pr(H_1) = 0.1, \Pr(H_0) = 0.9$).

Each panel from Figure 2 shows that Killeen's p_{rep} considerably overestimates the model–averaged p_{rep}^{Bma}. The extent of overestimation increases with n. Figure 2 highlights two causes that lead Killeen's p_{rep} to overestimate the probability of concurrence. First, the difference between the solid and the dashed lines indicates the extent to which the data support H_0. When $f(D|H_0)$ is non–negligible compared to $f(D|H_1)$, the non–negligible posterior probability for H_0 drives down p_{rep}^{Bma} compared to p_{rep}. Second, the difference between the dashed and the dotted lines indicates the additional effect of prior plausibility—when H_1 is unlikely *a priori*, the probability of concurrence is low, especially for data that are inconclusive.

At this point, it is important to address two objections that might be raised against our analysis. The first objection holds that $\Pr(H_0) = 0$, because the null hypothesis is supposedly never true exactly. At first sight, this objection—should it be correct—appears to seriously undercut our analysis. The second objection holds that prior probabilities for models and parameters can never be known, and can therefore be safely ignored.

Objection 1: Is the Null Hypothesis Ever Exactly True?

The first objection to our analysis goes as follows (e.g., Cohen, 1994). We know that, when we compare two populations, the difference between them will never be exactly zero. This means that we should be able to demonstrate the existence of any effect whatsoever, given only that the sample size is large enough. For a frequentist hypothesis test, the argument states that for large n, we are guaranteed to reject the null hypothesis. But if we know beforehand that the null hypothesis should be rejected for large n, then why would we be interested in showing that it can or cannot be rejected for small n? For a Bayesian hypothesis test, the same argument could be used to claim that we know *a priori* that $\Pr(H_0) = 0$, as $\delta \neq 0$ always; and if this claim is true, Equation 3 simplifies to $p_{rep}^{Bma} = p_{rep}$, and the overestimation of p_{rep} is illusory.

The first counter–argument to this claim is that our results hold regardless of whether the null hypothesis is true exactly or only *approximately*. Specifically, the same qualitative pattern of results is obtained when we define the null hypothesis as a distribution of small effect sizes that is centered around zero. For mathematical details we refer the reader to Appendix B and to the work by Berger and Delampady (1987, pp. 321–322).

Intuitively, the idea is that if the null hypothesis is only true approximately, so that a very small effect is present—albeit one that would take a much larger sample size to detect reliably—then the probability of concurrence may not be exactly equal to 1/2, but it will be only slightly higher. For a null hypothesis that is only approximately true, the probability of concurrence might be, say, .51, and this value is much smaller than the values provided by Killeen's p_{rep}.

The second counter–argument is that the Bayes factor implements an automatic Ockam's razor that obeys the principle of parsimony (Myung & Pitt, 1997). This means that

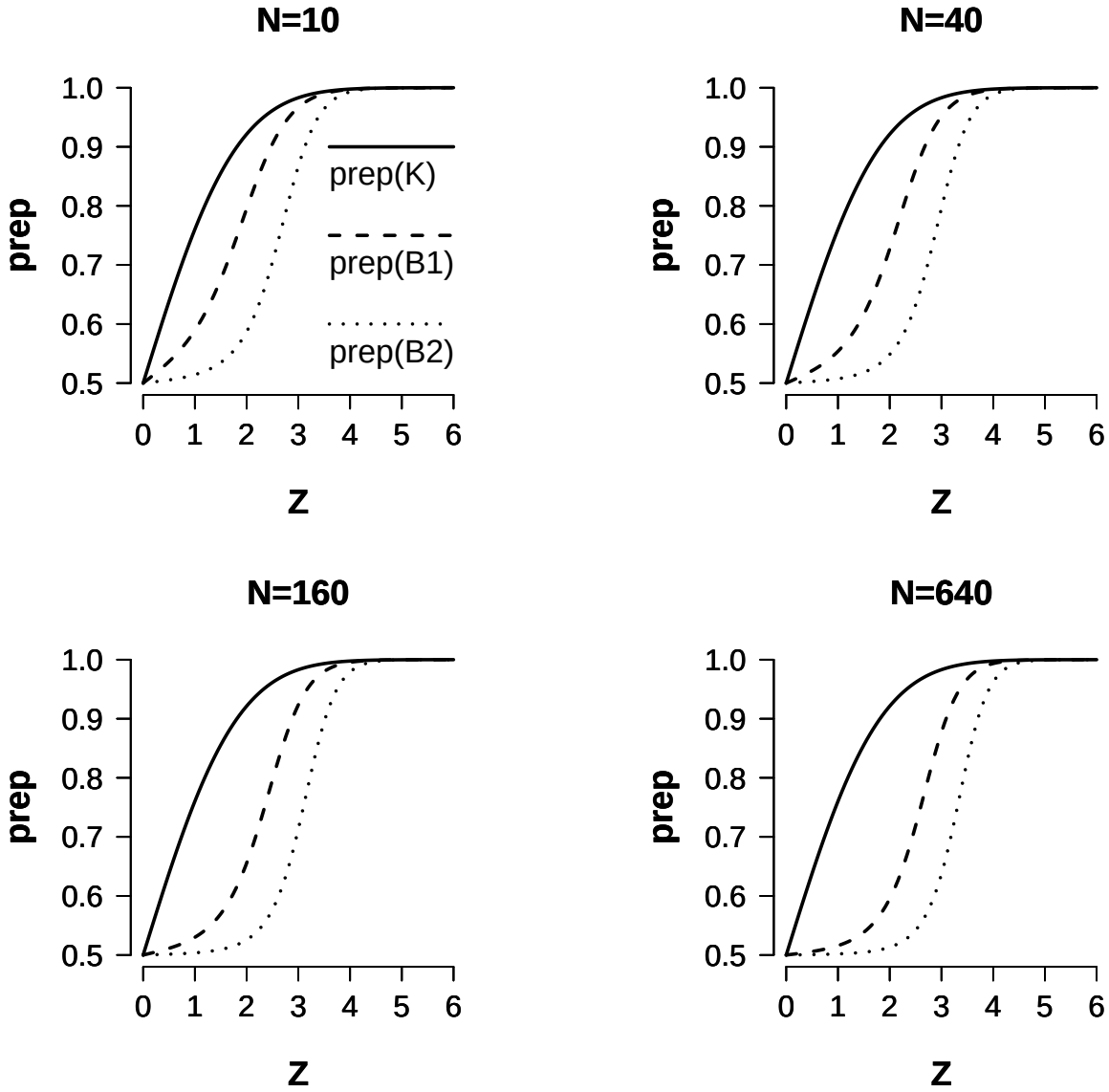

Figure 2. Results for three different p_{rep} methods as a function of Z–score. Prep(K) denotes the p_{rep} method proposed by Killeen (2005); Prep(B1) denotes a Bayesian model–averaged p_{rep}^{Bma} method with an H_1 that is *a priori* likely (i.e., $\Pr(H_1) = 0.5$); Prep(B2) denotes the same Bayesian model–averaged p_{rep}^{Bma} method but now with an H_1 that is a priori unlikely (i.e., $\Pr(H_1) = 0.1$).

the Bayes factor will prefer the model, of the two, that has the smallest one–step–ahead prediction error to unseen data from the same source (e.g., Dawid, 1984; Wagenmakers, Grünwald, & Steyvers, 2006). In other words, the Bayes factor prefers the model, of the two, that generalizes best. This predictive interpretation of the Bayes factor "does not depend on viewing one of the models as "true". (...) Thus the Bayes factor can be viewed as measuring the relative success of H_1 and H_0 at predicting the data." (Kass & Raftery, 1995, p. 777).

The third counter–argument to the claim that the null hypothesis is never true is that theories and models often predict the complete absence of an effect (Rouder, Speckman, Sun, Morey, & Iverson, in press). A test of these models therefore requires that we take the null hypothesis seriously. In the field of visual word recognition, for instance, the entry–opening theory (Forster, Mohan, & Hector, 2003) predicts that masked priming is absent for items that do not have a lexical representation; another example from that literature concerns the work by Bowers, Vigliocco, and Haan (1998), who have argued that priming effects are equally large for words that look the same in lower and upper case (e.g., kiss/KISS) or that look different (e.g., edge/EDGE), a finding supportive of the hypothesis that priming depends on abstract letter identities. A final example comes from the field of recognition memory, where "context noise" theories (e.g., Dennis & Humphreys, 2001), unlike rival "item noise" theories (e.g., Gillund & Shiffrin, 1984), predict the absence of a list–length effect. Empirically, this means context noise models predict no change in recognition performance for study lists of different lengths, and so their predictions are exactly those of the null hypothesis (Dennis, Lee, & Kinnell, 2008).

The above models do not predict that the experimental effects will be small; they predict them to be altogether absent. In fact, for theoretical purposes it often does not matter how large an effect is, as long as it is reliably detected. For instance, if priming effects were larger for words that look the same in lower and upper case (e.g., kiss/KISS) than for those that look different (e.g., edge/EDGE), this would undermine the hypothesis that letters are represented abstractly, no matter whether the effect size was 100 ms or 10 ms. Of course, it is much more difficult for a 10 ms effect to gain credence in the field, but this issue is orthogonal to the argument. Should the 10 ms effect be found repeatedly in different labs across the world, the effect would at some point be deemed reliable and considered strong evidence against any theoretical account that predicted its absence.

Finally, we believe that the philosophy that motivated the introduction of p_{rep} is consistent not with parameter estimation—in which the focus is generally on a single model—but rather with model selection:

> "(...) it is rare for psychologists to need estimates of parameters; we are more typically interested in whether a causal relation exists between independent and dependent variables (...). Are women attracted more to men with symmetric faces than to men with asymmetric faces? Does variation in irrelevant dimensions of stimuli affect judgments on relevant dimensions? Does review of traumatic events facilitate recovery?" (Killeen, 2005a, p. 345)

When we are interested in the question of "whether a causal relation exists between independent and dependent variables", we need to give special consideration to the possibility that such a relation is absent. Or, in the words of Jeffreys, "Any significance test whatever

involves the recognition that there is something special about the value 0, implying that the simple law [the null hypothesis H_0] may possibly be true" (Jeffreys, 1961, p. 394). Thus, one may certainly argue that the null hypothesis is always false, so that p_{rep}^{Bma} reduces to p_{rep}; but by doing so, one can no longer assess "whether a causal relation exists between independent and dependent variables", as the analysis would implicitly assume that such a relation is present.

In sum, we believe that the null hypothesis is not always false, and—perhaps more convincingly—our analysis does not critically depend on the absolute truth of H_0. Quantitatively and qualitatively similar results are obtained when we assume that H_0 is represented by a small distribution of effect sizes centered on zero (e.g., Appendix B; see also Berger & Delampady, 1987, pp. 321–322).

Objection 2: What About Those Priors?

Our analysis is Bayesian, and therefore involves priors, both on the level of models and on the level of parameters. For instance, we have discussed the effects of prior plausibility for H_0 and H_1 on the estimation of the probability of concurrence. Some researchers feel that priors are unknowable or at least subjective, and therefore have no place in scientific reasoning. When one dismisses the concept of priors, one dismisses the entire Bayesian approach, and, so it seems, our entire line of argumentation.

This objection is vulnerable to several counter–arguments. First, the fact that one is unable or unwilling to calculate a quantity does not mean that quantity is irrelevant and can be safely ignored. For instance, we may not know $\Pr(H_1)$, but the knowledge that it influences the probability of concurrence remains valuable. Second, as was outlined previously, the derivation of p_{rep} is Bayesian, and Killeen argued in fact that the crucial difference between p_{rep} and the p–value is that only the former is "a valid posterior predictive probability" (Killeen, 2005b, p. 1011). Third, we agree with Jim Berger, who argued that "(...) when different reasonable priors yield substantially different answers, can it be right to state that there *is* a single answer? Would it not be better to admit that there is scientific uncertainty, with the conclusion depending on prior beliefs?" (Berger, 1985, p. 125). Finally, the effect of priors can be formally assessed through *robustness analysis* (e.g., Berger, 1990).

More concretely, consider the priors involved in our analysis. The first prior is on the level of models, that is, $\Pr(H_0)$ and its complement, $\Pr(H_1)$. Equation 3 and Figure 2 highlight how these model priors impact on the probability of concurrence. Of course, the choice of a model prior may be highly subjective; a researcher who wants to demonstrate support for H_0 could assign it a relatively high prior probability, say $\Pr(H_0) = .9$. Similarly, a researcher who wants to demonstrate support for H_1 could bias the analysis in its favor by assigning H_1 a relatively high prior probability. A highly biased researcher would even go as far as to assign H_1 all prior probability: $\Pr(H_1) = 1$ and $\Pr(H_0) = 0$, which in fact yields the p_{rep} measure that is currently standard. In Bayesian model selection, one often avoids such *a priori* biases by equating the prior model probabilities, such that $\Pr(H_1) = \Pr(H_0) = 1/2$. These uninformative model priors seem appropriate in the absence of strong prior knowledge, because they "provide the level playing field necessary for unbiased evaluation." (Killeen, 2005b, p. 1011).

The second prior is on the level of parameters. Specifically, the standard calculation of p_{rep} assumes a uniform prior on effect size, that is, a normal prior $N(0, \tau^2)$ with $\tau \to \infty$. This prior is "improper" and is guaranteed to reduce the posterior model weight for H_1 to zero, regardless of what the data show (i.e., $\Pr(H_1|D) \to 0$ as $\tau \to \infty$, e.g., Kass & Raftery, 1995)—hence our reliance on the Smith and Spiegelhalter Z–test (i.e., Equation 6). One alternative is to use prior knowledge to fix τ to a reasonable number. For instance, based on a review of 474 research literatures in social psychology (Richard, Bond, & Stokes-Zoota, 2003), Killeen (2007) reported that the distribution of effect sizes was approximately Normal with variance 0.3. Denoting the hypothesis for which $\tau \ll \infty$ by H_1', one might specify $H_1' : \delta \sim N(0, \tau^2), \tau = \sqrt{0.3}$, and calculate both the probability of concurrence and the model weights from a $N(0, 0.3)$ prior on effect size.

Another attractive alternative is to carry out a robustness analysis[5]; this means that one computes p_{rep}^{Bma} for many different values of τ, every time using the $N(0, \tau^2)$ prior to calculate both the probability of concurrence and the model weights. That is, we consider two models, $H_0 : \delta = 0$ and $H_1'(\tau) : \delta \sim N(0, \tau^2)$, and use Equation 3 to calculate $p_{rep}^{Bma}(\tau)$ for many values of τ. Interest may then center, for example, on the maximum value for $p_{rep}^{Bma}(\tau)$ that can be attained by varying τ. We carried out such an analysis and confirmed that even the maximum value of $p_{rep}^{Bma}(\tau)$ is substantially lower than p_{rep}. For instance, our earlier numerical example referred to a hypothetical experiment that yields $d = 0.56$ with $n = 25$, resulting in $Z = 1.98$. For these data, the traditional p_{rep} equals .919. In sharp contrast, a robustness analysis revealed that the upper bound on $p_{rep}^{Bma}(\tau)$ is .764. This upper bound was obtained by varying the prior on effect size (i.e., τ in $\delta \sim N(0, \tau^2)$), and therefore does not depend on the possibly subjective choice for any specific parameter prior. The only prior that influences this result is the uninformative model prior that deems both H_0 and H_1' equally likely *a priori*. Consistent with our analysis using the Spiegelhalter and Smith Z–test, the robustness analysis supports our claim that p_{rep} overestimates the probability of concurrence.

General Discussion

In this article, we have shown that Killeen's p_{rep} statistic overestimates the probability of finding a concurrent result in a replicate experiment. The reason for the exaggeration is that p_{rep} assumes that the null hypothesis—under which the probability of concurrence is 1/2—can be ignored. We introduced a Bayesian model averaging procedure to show how the presence of a plausible null hypothesis tempers the enthusiasm stemming from Killeen's p_{rep}. When the data do not decisively rule out the null hypothesis, or when the null hypothesis is *a priori* much more likely than the alternative hypothesis (Macdonald, 2005), the probability of concurrence can be considerably lower than is advertised by p_{rep}.

Some of our Bayesian reasoning can also be brought to bear against the traditional p–value. A p–value indicates the probability of encountering a test statistic at least as extreme as the one that was observed in the experiment, given that the null hypothesis is true. This means that a p–value refers to the probability of data given the null hypothesis, and does not refer to the probability of the null hypothesis given data. When researchers study unlikely, counterintuitive phenomena (such as those commonly reported in high–impact

[5]We thank an anonymous reviewer for bringing this to our attention.

psychology journals), the statistical result "$p = .04$" does not warrant the conclusion that "the null hypothesis can be rejected". When one studies ESP, for instance, it takes more than $p = .04$ to reject the null hypothesis. From a Bayesian perspective, extraordinary claims require extraordinary evidence. It is important to note, however, that our argument against p_{rep} holds whenever H_0 is assigned any prior mass, and becomes more compelling as $\Pr(H_0) \rightarrow \Pr(H_1)$, leading to the uninformative model prior that provides "the level playing field necessary for unbiased evaluation." (Killeen, 2005b, p. 1011). The extent to which p_{rep} overestimates the probability of concurrence when $\Pr(H_0) = \Pr(H_1)$ can be seen by comparing Prep(K) to Prep(B1) in Figure 2.

Although it may be argued that our analysis is restricted to the framework of Bayesian inference, this does not diminish its relevance for the evaluation of p_{rep}. Doros and Geier (2005) have argued that p_{rep} is exclusively a Bayesian concept, and p_{rep} should therefore be susceptible to Bayesian arguments. Also, Equation 3, Appendix B, and our robustness analysis indicate that our general conclusion (i.e., p_{rep} overestimates the probability of concurrence) holds across all model priors and across a large class of parameter priors. It is possible to construct a parameter prior under which our general conclusion does not hold; such a parameter prior would be asymmetrical around zero, and assign a lot of mass to values greater or smaller than zero. In the absence of strong prior knowledge, we do not feel such priors are appropriate for Bayesian hypothesis testing, a sentiment that is echoed by the absence of such priors in the Bayesian literature.

Practical Implications

Consider again the hypothetical situation in which you conduct an experiment to test whether words such as *pizza* prime words such as *coin* (i.e., perceptual priming, Pecher et al., 1998). Recall that the experimental effect yields $p_{rep} \approx .94$. What should we make of this value? We hazard to guess that researchers, reviewers, and editors are likely to (mis)interpret $p_{rep} \approx .94$ as follows: "If this experiment were to be repeated, there is a 94% chance to again observe a reliable perceptual priming effect. This is strong evidence for the presence of perceptual priming in the original experiment."

Unfortunately, this interpretation is as tempting as it is wrong. First, "replication" does not refer to finding again a result that is reliable or statistically significant; in the context of p_{rep}, "replication" refers to concurrence, that is, finding again a result of the same sign, however small and insignificant. This means that the lowest possible value for p_{rep} is already as high as 0.5. It is debatable whether researchers are interested in the probability of concurrence rather than the probability of replication in the traditional sense (i.e, the probability of a replication experiment again yielding a significant result). Another problem with "concurrence" is that it is a definition of replication that is difficult to gauge; how impressive is it that the probability of concurrence is .85, .95, or .99?

Second, our work here shows that p_{rep} is a valid estimate of concurrence only when the null hypothesis can be completely ruled out and when the alternative hypothesis holds that all effect sizes are equally likely *a priori*. Together, this means that p_{rep} does not estimate the probability of concurrence, but that it estimates an upper bound for this probability, a bound that holds only under strict and arguably unrealistic assumptions. Thus, the correct interpretation of $p_{rep} \approx .94$ is "the chance of a concurrent result in a replication experiment is likely to be lower than .94". Although this statement is correct, it does not appear to

provide much insight.

In general, we believe that widespread adoption of p_{rep} can all too easily mislead researchers into thinking that their effects are more reliable than they really are. In a field such as psychology, where there is pressure to publish and replication research is relatively rare (Lindsay & Ehrenberg, 1993), this means that the p_{rep} statistic may unwittingly facilitate the dissemination of Type I errors, that is, findings that do *not* replicate.

So what options are we left with? Researchers who believe that concurrence is a meaningful concept may replace p_{rep} with a model averaged version such as p_{rep}^{Bma}. Researchers who are skeptical about the very idea of concurrence may resort to alternative methods for statistical inference, a discussion of which is the topic of the next section.

The Future of Statistical Inference in Psychology

For many decades, researchers have pointed out the many shortcomings of p–value hypothesis testing (e.g., Cohen, 1994; Edwards, Lindman, & Savage, 1963; Wagenmakers, 2007). The p_{rep} statistic was developed to address some of these shortcomings, but, unfortunately, it is sensitive to some shortcomings of its own. This raises the question of whether there is a single method for statistical inference method that has no shortcomings at all. The answer, alas, appears to be in the negative. Hypothesis testing is very difficult. Nickerson (2000, p. 290) summarized the situation as follows: "NHST [Null Hypothesis Statistical Testing—IWL] surely has warts, but so do all the alternatives". A pragmatic solution would be to use more than just a single method for inference, and demonstrate that the conclusions hold regardless of the particular method that is used.

What are the alternatives to p–values and p_{rep}? They include Bayesian procedures (e.g., Hoijtink, Klugkist, & Boelen, 2008; Kass & Raftery, 1995; Klugkist et al., 2005; Lee & Wagenmakers, 2005; Rouder et al., in press; Raftery, 1995; Wagenmakers, 2007), Bayesian–frequentist compromises (e.g., Berger, 2003; Berger, Boukai, & Wang, 1997; Berger, Brown, & Wolpert, 1994; Good, 1983), Akaike's Information Criterion (AIC; e.g., Akaike, 1974; Burnham & Anderson, 2002), cross–validation (e.g., Browne, 2000; Geisser, 1975; Stone, 1974), bootstrap methods (e.g., Efron & Tibshirani, 1997), prequential methods (e.g., Dawid, 1984; Wagenmakers et al., 2006) and methods based on the principle of Minimum Description Length (MDL; e.g., Grünwald, 2000; Grünwald, Myung, & Pitt, 2005; Pitt, Myung, & Zhang, 2002; Rissanen, 2001). All these methods are methods for *model selection*, in that the explicit or implicit goal is to compare different models and select the best one (for applications of model selection in the field of psychology see the two special issues in the *Journal of Mathematical Psychology*: Myung, Forster, & Browne, 2000; Wagenmakers & Waldorp, 2006). Methods for model selection do not assess the adequacy of H_0 or H_1 in isolation. Rather, the adequacy of H_0 is compared to the adequacy of an alternative model, H_1, automatically avoiding the negative consequences that arise when the focus is on a single model.

In experimental psychology, model selection procedures are mostly used to adjudicate between nonnested, complicated nonlinear models of human cognition. There is no reason, however, why these procedures could not be applied to run–of–the–mill statistical inference problems involving nested linear models such as ANOVA (Lee & Pope, 2006). We hope and expect that in the near future, concrete alternatives to p–values (e.g., Bayesian hypothesis tests) will be developed and made available in a way that benefits the majority of exper-

imental psychologists. This is an exciting possibility that could change the landscape of statistical inference in psychology in a fundamental way.

In conclusion, we applaud Killeen's effort to have psychological researchers compute a Bayesian quantity to decide whether or not there is a causal relation between independent and dependent variables. Unfortunately, the choice for p_{rep} is beset by serious problems (e.g., Iverson et al., in press a, in press b), one of which is that it can lead to overconfidence and undue optimism. We recommend that researchers do not report p_{rep} but either report a model–averaged version of p_{rep} or report the conclusions from one or more alternative methods of statistical inference.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Ashby, F. G., & O'Brien, J. B. (2008). The p_{rep} statistic as a measure of confidence in model fitting. *Psychonomic Bulletin & Review*, *15*, 16–27.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer.

Berger, J. O. (1990). Robust Bayesian analysis: Sensitivity to the prior. *Journal of Statistical Planning and Inference*, *25*, 303–328.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18*, 1–32.

Berger, J. O., Boukai, B., & Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statistical Science*, *12*, 133–160.

Berger, J. O., Brown, L., & Wolpert, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. *The Annals of Statistics*, *22*, 1787–1807.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*, 109–122.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory.* New York: Wiley.

Bowers, J. S., Vigliocco, G., & Haan, R. (1998). Orthographic, phonological, and articulatory contributions to masked letter and word priming. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1705–1719.

Browne, M. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*, 108-132.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information–theoretic approach (2nd ed.).* New York: Springer Verlag.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Congdon, P. (2003). *Applied Bayesian modelling.* Chichester, UK: Wiley.

Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, *16*, 1002–1004.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, *147*, 278–292.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*, 452–477.

Dennis, S. J., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list–length effect. *Journal of Memory and Language*, *59*, 361–376.

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

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B*, *57*, 45–97.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Efron, B., & Tibshirani, R. (1997). Improvements on cross–validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, *92*, 548–560.

Forster, K. I., Mohan, K., & Hector, J. (2003). The mechanics of masked priming. In S. Kinoshita & S. J. Lupker (Eds.), *Masked priming: The state of the art* (pp. 3–38). New York, NY: Psychology Press.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference.* Boca Raton, FL: Chapman & Hall/CRC.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*, 320–328.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach.* Boca Raton (FL): CRC Press.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1–67.

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications.* Minneapolis: University of Minnesota Press.

Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 2* (pp. 249–269). New York: Elsevier.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*, 133–152.

Grünwald, P., Myung, I. J., & Pitt, M. A. (Eds.). (2005). *Advances in minimum description length: Theory and applications.* Cambridge, MA: MIT Press.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382-417.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses that are of practical value for social scientists.* New York: Springer.

Iverson, G., Lee, M. D., & Wagenmakers, E.-J. (in press b). p_{rep} misestimates the probability of replication. *Psychonomic Bulletin & Review.*

Iverson, G., Lee, M. D., Zhang, S., & Wagenmakers, E.-J. (in press a). p_{rep}: An agony in five fits. *Journal of Mathematical Psychology.*

Jeffreys, H. (1961). *Theory of probability.* Oxford, UK: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 377–395.

Killeen, P. (2005). Tea–tests. *The General Psychologist*, *40*, 15–18.

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

Killeen, P. R. (2007). Replication statistics as a replacement for significance testing: Best practices in scientific decision–making. In J. W. Osborne (Ed.), *Best practices in quantitative methods.* Thousand Oaks, CA: Sage Publications.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477–493.

Lee, M. D., & Pope, K. J. (2006). Model selection for the rate problem: A comparison of significance testing, Bayesian, and Minimum Description Length statistical inference. *Journal of Mathematical Psychology*, *50*, 193–202.

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, *49*, 293–337.

Lindsay, R. M., & Ehrenberg, A. S. C. (1993). The design of replicated studies. *The American Statistician*, *47*, 217–228.

Macdonald, R. R. (2005). Why replication probabilities depend on prior probability distributions: A rejoinder to Killeen (2005). *Psychological Science*, *16*, 1007–1008.

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*, 1535–1546.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection [Special issue]. *Journal of Mathematical Psychology*, *44* (1–2).

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society B*, *57*, 99–138.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.)*. London: Arnold.

Pecher, D., Zeelenberg, R., & Raaijmakers, J. G. W. (1998). Does pizza prime coin? Perceptual priming in lexical decision and pronunciation. *Journal of Memory and Language*, *38*, 401–418.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Cambridge: Blackwells.

Richard, F. D., Bond, C. F. J., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, *47*, 1712–1717.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (in press). Bayesian t–tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*.

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*, 471-481.

Smith, A. F. M., & Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society B*, *42*, 213–220.

Stone, M. (1974). Cross–validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B*, *36*, 111–147.

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

Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing. *Psychological Science*, *17*, 641–642.

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, *50*, 149–166.

Wagenmakers, E.-J., & Waldorp, L. (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, *50*(2).

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

Appendix A
Bayesian Results for p_{rep}

Assume that one has a normal prior $N(0, \tau^2)$ on the effect parameter $\delta = \frac{\mu_E - \mu_C}{\sigma}$, where subscripts E and C refer to an experimental and control group, respectively. The equation for Killeen's p_{rep} is recovered in the limit as $\tau \to \infty$.

It is convenient to introduce $\omega^2 = \frac{n}{2}\tau^2$ and $\theta = \frac{\omega^2}{1+\omega^2}$. Here, n is the common sample size of the experimental and control groups. Note that $0 < \theta < 1$ and $\lim_{\omega\to\infty}\theta = 1$, and that ω is large for large τ and for large n.

Suppose one has an observed effect d at hand. The posterior for δ is $\delta|d \sim N(d\theta, \theta 2/n)$, and—assuming a replication experiment employs the same sample size n per group as the original—the posterior predictive density for d_{rep} is $d_{rep}|d \sim N(d\theta, (1+\theta)2/n)$. We get

$$p_{rep}|d, \tau, n = \Pr(d_{rep}d \geq 0|d, \tau, n) = \Phi\left[\theta\frac{|d|\sqrt{n/2}}{\sqrt{1+\theta}}\right]. \tag{8}$$

It is easy to check that

$$p_{rep}|d, \tau, n < \lim_{\tau\to\infty} p_{rep}|d, \tau, n = p_{rep} = \Phi\left[\frac{|d|\sqrt{n/2}}{\sqrt{2}}\right], \tag{9}$$

where the last expression was given by Killeen and is the one recommended by *Psychological Science*.

It is also convenient to write $Z = d\sqrt{n/2}$ (note that Z is the familiar frequentist test statistic), and derive the more compact equation

$$p_{rep}|d, \tau, n = \Phi\left[\frac{|Z|\theta}{\sqrt{1+\theta}}\right]. \tag{10}$$

The usual expression for p_{rep} is obtained when $\theta = 1$.

Appendix B
Generality of our Result

Equation 3 is subject to the distraction that some people believe that $\Pr(H_0|d) = 0$. Note however that the inequality from Equation 3 is far more general than indicated. Suppose we replace $H_0 : \delta = 0$ by a model M_0 that involves a prior $N(0, \epsilon^2)$ on δ, and we contemplate another competing model M_1' which involves a prior $N(0, \tau^2)$ on δ (replacing the flat improper prior). Assume that $\epsilon^2 < \tau^2$ (typically, $\epsilon^2 \ll \tau^2$).

Write $\zeta = \frac{\epsilon^2 n/2}{1+\epsilon^2 n/2}$ and $\theta = \frac{\tau^2 n/2}{1+\tau^2 n/2}$. Then we get

$$p_{rep}^{\zeta} \triangleq \Pr\left(d_{rep}d \geq 0|d, M_0\right) = \Phi\left(\zeta\frac{|d|\sqrt{n/2}}{\sqrt{1+\zeta}}\right), \tag{11}$$

and

$$p_{rep}^{\theta} \triangleq \Pr\left(d_{rep}d \geq 0|d, M_1'\right) = \Phi\left(\theta\frac{|d|\sqrt{n/2}}{\sqrt{1+\theta}}\right), \tag{12}$$

where $\triangleq$ means "by definition".

In these terms we have

$$\begin{aligned} p_{rep}^{Bma} = \Pr\left(d_{rep}d \geq 0|d\right) &= \Pr\left(M_0|d\right)\Pr\left(d_{rep}d \geq 0|d, M_0\right) + \Pr\left(M_1'|d\right)\Pr\left(d_{rep}d \geq 0|d, M_1'\right) \\ &= \Pr\left(M_0|d\right) \times p_{rep}^{\zeta} + \Pr\left(M_1'|d\right) \times p_{rep}^{\theta} \\ &< p_{rep}^{\theta} \text{ [since } \epsilon^2 < \tau^2\text{, we have } \zeta < \theta \text{ and } p_{rep}^{\zeta} < p_{rep}^{\theta}] \\ &< p_{rep} \text{ [since } p_{rep}^{\theta} \text{ is increasing in } \theta \text{ and } p_{rep} = \lim_{\theta \to 1} p_{rep}^{\theta}] \end{aligned} \tag{13}$$

So we see that our point does *not* rely on the assumption that $\Pr(H_0) \neq 0$.

To summarize: Under any assumption on ζ, θ subject to $\zeta < \theta$ and any assumption on the prior probability $\Pr(M_0)$ it is the case that $p_{rep}^{Bma} < p_{rep}$. Note that the difference $p_{rep} - p_{rep}^{Bma}$ is calculable given the values of ϵ^2, τ^2, n, and d.

It is now easy to argue that p_{rep} can and often does overstate the evidence provided by d that a faithful replication will agree in direction with d. Only people who truly believe that M_0 is impossible, yet are so uncertain as to the range of observed effects that they adopt a flat improper prior, would expect that $p_{rep}^{Bma} = p_{rep}$. Such people know so much and yet so little.

Of course what one would really like is a useful lower bound so that one could report that, on the basis of given data, the probability of achieving agreement in a replication as to direction is at least that lower bound. Alas the only general lower bound seems to be 1/2 and that is not very interesting.