Running head: EVIDENCE FOR FEELING THE FUTURE

# An Assessment of The Evidence For Feeling The Future With A Discussion of Bayes Factor and Significance Testing

Jeffrey N. Rouder

University of Missouri

Richard D. Morey

University of Groningen

## Abstract

We provide a statistical assessment of the claim that people can feel the future from the results provided in Bem (in press). Conventional significance testing is particularly ill-suited for assessing the evidence for competing positions because it may not be used to state evidence for the null hypothesis, and, consequently, is biased to overstate the evidence against it. Bayes factors, in contrast, serve as well-calibrated measures of relative evidence for two competing theoretical position, which, in this case, are that feeling-the-future hypothesis holds or does not. Bayes factors describe how researchers should update their prior beliefs about the odds of hypotheses in light of data. We find the evidence that people can feel the future with neutral and erotic stimuli to be slight, with Bayes factors of 3.23 and 1.57, respectively. There is, however, a surprising degree of evidence for the hypothesis that people can feel the future with emotionally-valenced nonerotic stimuli, with a Bayes factor of about 40. Though this value is certainly noteworthy, it is several orders of magnitude lower than what is required to overcome appropriate skepticism of such implausible claims. Moreover, this value serves as an upper bound on evidence as no corrections are made for the exploratory nature of the comparisons or the possible unreported studies that fail to find feeling-the-future effects.

# An Assessment of The Evidence For Feeling The Future With A Discussion of Bayes Factor and Significance Testing

*"Although our colleagues in other disciplines would probably agree with the oft-quoted dictum that extraordinary claims require extraordinary evidence, we psychologists are more likely to be familiar with the methodological and statistical requirements for sustaining such claims...."* (Bem, in press).

Bem (in press) concludes that people can feel or sense salient events in the future that could not otherwise be anticipated. For example, in his Experiment 2, Bem presents participants with two rather ordinary pictures, and asks them to indicate which one will be subsequently chosen by a random number generator. If a participant correctly anticipated the random choice, she or he was subsequently rewarded with a brief display of positively-valenced picture. Conversely, if a participant incorrectly anticipated the random choice, he or she was subsequently punished with a negatively-valenced picture. Bem claimed that people could indeed feel these future reward and punishment events, and consequently, were able to anticipate the random choice at a rate significantly better than chance. Bem presents a sequence of similar experiments and results, and on this basis concludes that people can feel the future. This phenomenon and others like it in which people can show seemingly impossible awareness of events are termed psi phenomena.

If psi phenomena are substantiated, they would be among the most important findings in the history of psychology. The existence of psi phenomena would force us to revise not only our theories of psychology, but of biology and physics as well. In our view, when seemingly implausible claims are made with conventional methods, it provides an ideal moment to re-examine these methods. Our focus here is on the statistical inference in Bem's paper: we ask, *"How shall we evaluate evidence provided by data?"* The conventional approach used by Bem is significance testing, in which the analyst evaluates a $p$-value against a criterion. We highlight a little-appreciated critique that inference by $p$-values overstates the evidence against the null (Edwards, Lindman, & Savage, 1963). In the current case, the null serves as the reasonable statement that there is no psi phenomena, and it is a disservice to the field to overstate the evidence against it. Wagenmakers et al. (submitted) offers the same critique, and then assesses the evidence in each of Bem's nine experiments. In this paper, we provide a meta-analytic assessment of the total evidence in Bem's report. To foreshadow, our assessment is that Bem's experiments, collectively, provide some evidence of psi phenomena, but not enough to sway the beliefs an appropriately skeptical reader.

## The evidence from $p$-values

There is a well-known asymmetry in significance testing: researchers can reject the null hypothesis, but can never accept it. This asymmetry works against the goals of scientific inquiry because null hypotheses often correspond to theoretically useful statements of invariance and constraint (Gallistel, 2009; Kass, 1992; Rouder, Speckman, Sun, Morey, & Iverson, 2009). For Bem's case, the null hypothesis is the theoretically attractive, reasonable, and highly interpretable constraint that there are no psi phenomena. In order to fairly assess the evidence about psi, it is necessary to be able to state the evidence for or against the null provided by the data. Should the null hypothesis of no psi hold, researchers must be able to state evidence for it. Especially in the case of psi research, the null hypothesis is not simply a straw man that deserves only to be rejected, yet this is exactly how it is treated by significance testing.

The above point about asymmetry is easy to grasp. Its implications, however, are subtle and consequential as they extend beyond not being able to state evidence for the null hypothesis; they extend to assessing evidence in the data for the alternative as well. A good starting point is consideration of the distribution of $p$-values under two competing hypotheses; examples are shown in Figure 1A. If the null hypothesis is false, then $p$-values tend to be small, and decrease as sample sizes increase. The dashed line colored green shows the distribution of $p$-values when the underlying effect size is .2 and the sample size is 50; the dashed-dotted line colored red shows the same when the sample size is increased to 500. The distribution of $p$-values under the null, however, is quite different. Under the null, all $p$-values are equally likely (solid line colored blue in Figure 1A). Perhaps surprisingly, this distribution holds regardless of sample size; $p$ values do not increase under the null as sample sizes increase.

The logic behind significance testing is a form of argument by contradiction. If data are improbable under the null, then the null is contradicted, and presumably, there is some alternative under which the data are more probable. We ask about the factor by which the data are more probable under the alternative than under the null. This factor serves as a measure of evidence for the alternative relative to the null. Suppose a data set with sample size of 50 yields a $p$-value in the interval between .04 and .05. Figure 1B shows the distributions of $p$-values for the null and the alternative (effect size = .2) around this interval, and the probabilities are the shaded areas under the curve. The probability of observing a $p$-value under the null and alternative is .01 and .04, respectively. Therefore, the alternative fares four times better than the null. Although such a ratio constitutes evidence for the alternative, it is not as substantial as might be inferred by such a small $p$-value.

Figure 1C shows a similar plot for the null and alternative (effect size = .2) for a large sample size of 500. For this effect size and sample size, very small $p$-values are the norm. In fact, a $p$-value between .04 and .05 is about ten times more likely under the null than under the alternative. More generally, a $p$-value at any one point, say .05, constitutes increasing evidence for the null in the large sample-size limit. This paradoxical behavior of significance testing in which researchers reject the null even though the evidence overwhelmingly favors it is known as *Lindley's Paradox* (Lindley, 1957), and is a primary critique of inference by $p$-values in the statistical literature.

We can examine the evidence from Bem's data for various alternatives relative to the null. In Experiment 1, for example, participants needed to anticipate which of two erotic pictures they would be shown. The average performance across 100 naive subjects was .531, and this level is significantly different from the at-chance baseline of .5 ($t(99) = 2.51$, $p = .007$). Figure 1D shows the evidence for various alternatives. The probability ratios on the y-axis are the probability of the observed $p$-value under a specific alternative relative to that under the null. Not surprisingly, these ratios vary greatly with the choice of alternative. Alternatives that are very near the null of .5, say, .525, are preferred over the null (filled circle in Figure 1D). Alternatives further from .5, say .58 (filled square) are definitely not preferred over the null. Note that even though the null is rejected at $p = .007$, there is only a small range of alternatives where the probability ratio exceeds 10, and for no alternative does it exceed 25, much less 100 (as might naively be inferred by the $p$-value). We see that the null may be rejected by $p$-values even when the evidence for every specific point alternative is more modest.

## Bayes factor measures of evidence

The probability ratio in Figure 1D may be denoted by $B$, and expressed as follows:

$$B = \frac{Pr(\text{data} \mid H_1)}{Pr(\text{data} \mid H_0)},$$

where $H_0$ is the null and $H_1$ is that the alternative is that true performance is a specific value, for example .52. In Bayesian statistics, probability ratios $B$ are called *Bayes factors*, and they are well-calibrated measures of evidence from the data for one hypothesis relative to another. One drawback of the preceding formulation, however, is that the alternative is a single point hypothesis. In Bayesian statistics it is possible and desirable to consider hypotheses in which parameters range over many possible values. To do this, the analysts specifies how each parameter value should be weighted. Figure 1D shows such weights (dashed lines), and for this alternative hypothesis, small effects are weighted more than large ones. The distribution of weights over parameters is called the *prior distribution*. When an alternative consists of a weighted range of parameter values, the probability of the data is

$$Pr(\text{data}|H_1) = \int Pr(\text{data}|\theta)f(\theta)d\theta,$$

where $\theta$ are the parameters, and $f$ is the prior distribution on these parameters. The probability of the data given the hypothesis is the expected or weighted averaged probability across the possible parameter values. The Bayes factor is

$$B = \frac{Pr(\text{data}|H_1)}{Pr(\text{data}|H_0)} = \frac{\int Pr(\text{data}|\theta)f(\theta)d\theta}{Pr(\text{data}|\theta = \theta_0)},$$

where $\theta_0$ is the value of $\theta$ under the null, or .5 for Figure 1D. For the shown prior, the Bayes factor evidence for the observed $p$-value is 3.23, or that the observed level of performance favors the alternative by a ratio a bit more than 3-to-1. This value differs from than in Wagenmakers et al. (submitted) because we used a one-sided prior weights (performance must be at or above chance) consistent with Bem's analysis. Wagenmakers et al., in contrast, chose a two-sided prior to express the position that assumptions about the direction of psi may be unwarranted.

Researchers must choose the prior distribution $f$. Fortunately, there is ample guidance in the literature about how to do so (e.g., Liang et al., 2008; Rouder et al., 2009; Wagenmakers, 2007). We opt for a set of default priors first proposed by Jeffreys (1961), and developed by Liang, Paulo, Molina, Clyde, and Berger (2008), Rouder et al. (2009), and Zellner and Siow (1980), among others. Rouder et al. (2009) and Wagenmakers, Lodewyckx, Kuriyal, and Grassman (2010) discuss this prior within the psychological literature, and provide easy-to-use web-based applets for computation.[1]

Bayes factor evidence is the probability ratio of data given hypotheses. A related quantity of interest is the probability ratio of hypotheses given data, called the *posterior odds.* The posterior odds describe how the analysts degree of belief in the hypotheses after observing the data. The following equation describes the relationship between posterior odds and Bayes factor:

$$\frac{Pr(H_1|\text{data})}{Pr(H_0|\text{data})} = B \times \frac{Pr(H_1)}{Pr(H_0)},$$

where the terms $\frac{Pr(H_1|\text{data})}{Pr(H_0|\text{data})}$ and $\frac{Pr(H_1)}{Pr(H_0)}$ are posterior and prior odds, respectively. The prior odds describe the beliefs about the hypotheses before observing the data. The Bayes factor describes how the evidence from the data should affect beliefs. For example, suppose the evidence from a set of psi-phenomena experiments yielded a Bayes factor of 40 in favor of psi. Consider the a skeptical reader with prior odds of million-to-one against psi. In this case, the reader should revise their beliefs by a factor of 40, to 25,000-to-one against psi. Likewise, a reader that has prior odds favoring psi should multiply these priors by 40 in light of the data to reach an even more favorable posterior odds. Bayes factors are logically independent of prior odds, and, consequently, are ideal for scientific communication (Jeffreys, 1961). We recommend that researchers report Bayes factors, and that readers use the context of prior knowledge, such as knowledge about physical laws or plausible mechanisms, to set prior odds in interpreting these Bayes factor.

## Evidence in Bem's Data

At the end of his paper, Bem requests that psychologists update their posterior beliefs about the psi phenomena in light of his data. The conventional methods used by Bem, however, do not provide a formal means of updating. Fortunately, the Bayes factor method advocated here is the optimal method of updating beliefs in light of data (Cox, 1946). Wagenmakers et al. (submitted) provide the Bayes factor for each of Bem's experiments, and these serve as the appropriate measures of evidence for each experiment taken individually. We provide a meta-analytic Bayes factor of the evidence across Bem's nine reported experiments, and this Bayes factor provides an appropriate summary measure of the total evidence. This meta-analytic approach has two main advantages: First, it allows us to gather strength across similar findings. Second, and perhaps more importantly, it allows us to weigh or compare the conflicting findings in Bem (2010). An example of such conflicting findings is the contrast between the positive findings of psi for emotionally valenced items in Bem's Experiments 2, 3, and 4 vs. the failure to observe psi for the same class of items in Experiment 1.

Bem's data afford the opportunity to separately assess whether people can feel the future for erotic events, emotional but not erotic events, and emotionally neutral events. The computations for the meta-analytic Bayes factor are provided in the Appendix, and the resulting values are shown in Table 1. We have not included results from Experiments #5, #6, and #7 in our meta-analysis because we are unconvinced that these are interpretable. These three experiments are retroactive mere-exposure effect experiments in which the influence of future events purportedly affects the current preference for items. The main difficulty in interpreting these experiments is forming an expectation about the direction of an effect, and this difficulty has consequential ramifications. In the vast majority of conventional mere-exposure effect studies, participants prefer previously viewed stimuli (Bornstein, 1989). Yet, Bem observes the opposite effect, habituation, which may be interpreted either as evidence for or evidence against psi. What is sorely needed is a conventional mere-exposure effect with the same stimuli to firmly establish expectations. In fact, Bem does just this with his retroactive priming experiments, and the inclusion of conventional priming studies to establish firm expectations greatly increases the interpretability of the results. Without these control experiments for mere exposure, the most judicious course is to ignore Experiments 5, 6, and 7 and rely of the remaining 6 experiments for assessment.

Table 1 reveals that there is relatively little support for the propositions that people can feel the future with erotic or neutral events. The Bayes factor does offer some support for a retroactive effect of emotionally-valenced, nonerotic stimuli: the evidence for an effect provided by Experiments #2, #3, and #4 outweighs the evidence against an effect provided by Experiment #1. In Experiment #2, participants were rewarded with brief presentations positive pictures and punished with brief presentations negative ones when they anticipated or failed to anticipate, respectively, the future state of a random-number generator. In Experiments #3 and #4, participants identified emotionally-valenced target stimulus more quickly when a subsequently presented prime matched the valence of the target.

## Discussion

Our Bayes factor analyses of Bem's data, which Bem offered as evidence of psi effects of erotic, neutral, and emotionally-valenced nonerotic stimuli, show that the data only support more modest claims. The data yield no substantial support for psi effects of erotic or neutral stimuli. For emotionally-valenced nonerotic stimuli, however, we found a Bayes factor of about 40, and this is the factor by which readers should increase their odds. We caution readers against interpreting this Bayes factor as the posterior odds that psi is true (Wagenmakers et al., submitted). On the contrary, posteriors odds should reflect the context provided by prior odds as discussed previously. As noted by Bem in the passage cited at the start of this article, psychologists are obligated to hold low prior odds for psi because these phenomena fundamentally challenge well-substantiated theories in psychology, physics, and biology without providing any plausible alternative mechanism. Hence, while the evidence provided by Bem is certainly worthy of notice, it should not be sufficient to sway an appropriately skeptical reader. We remain unconvinced of the viability of psi.

We have purposefully limited our focus to the statistical evaluation of evidence in Bem (in press). Wagenmakers et al. (submitted) discuss a host of issues including the *post hoc* nature of some contrasts, and the possibility of a so-called file drawer problem in which conditions that do not provide evidence for psi are not reported. In light of these other critiques, it is advisable to interpret the Bayes factor of about 40 as an upper bound on the evidence for psi in Bem's report.

Bem's controversial claims of feeling the future provide an opportunity to re-examine the methods by which psychological scientists assess evidence for positions. Our main methodological concern is that inference by $p$-values fails to seriously consider the null hypothesis as a viable possibility. Consequently, researchers who use it are apt to reject the null on the basis of insufficient evidence. We recommend that researchers adopt Bayes factor methodology, because this approach provides a rational and consistent assessment of the relative evidence between any two hypotheses (Edwards et al., 1963). Researchers who seek to adopt Bayes factor will benefit from the guidance and free applications provided in Gallistel (2009), Rouder et al. (2009), and Wagenmakers et al. (2010).

## Appendix: Meta-Analytic Bayes Factor

Computing a Bayes factor across several experiments appears to be straightforward at first glance. Readers may have the intuition that one should simply multiply odds. Unfortunately, this approach is not valid. As discussed by Rouder et al. (2009), Bayes factors respect the resolution of data: When the sample size is small, small effects may be considered evidence for the null as the null is the more parsimonious description given the resolution provided by the data. As the sample size grows however, the resolution provided for the data is finer, and small effects are more concordant with the alternative. An appropriate analogy may be a criminal court trial in which each of several witnesses provides only partial information as to the guilt of a defendant who has committed a crime. If the jury is forced to assess the odds after hearing the testimony of one witness, these odds may all favor innocence as no one witness is compelling enough in isolation to provide evidence for guilt. However, if the jury considers the totality of all testimonies, the weight will assuredly shift toward guilt.

Our approach is to consider two hypotheses for the sequence of experiments. The first, the null, is that the true effect size is zero for all experiments. The second is that there is a single true effect size greater than zero which underlies all the experiments. Consider a sequence of $t$-values, $t_1, t_2, \ldots t_M$, from $M$ comparisons. Under the null, the probability of this sequence is

$$P(t_1, \ldots t_M \mid H_0) = \prod_i^M g_t(t_i, N_i - 1),$$

where $g_t(t, \nu)$ is the probability density function of Student's $t$-distribution evaluated at $t$ with $\nu$ degrees of freedom. Under the alternative, the probability of a of this sequence is

$$P(t_1, \ldots t_M \mid H_1) = \int_{\delta=0}^{\delta=\infty} \prod_i^M h_t(t_i, N_i - 1, \sqrt{N_i}\delta) f(\delta)\, d\delta,$$

where $\delta$ is the effects size; $h_t(t, \nu, \lambda)$ is the probability density function of the noncentral $t$ distribution evaluated at $t$ with degrees of freedom $\nu$ and noncentrality parameter $\lambda$; and

$f$ is the prior weights on parameter $\delta$, the true effect size. We place a default one-tailed prior on $\delta$, namely that it follows a $t$-distribution with a single degree of freedom (Jeffreys, 1961; Rouder et al., 2009). For a single $t$ value, this Bayes factor is equivalent to the Bayes factor suggested by Rouder et al.

## References

Bem, D. (in press). Feeling the future: Experimental evidence for anamalous retroactive infleces on cognition and affect. *Journal of Personality and Social Psychology.*

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, *106*, 265–289.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, *14*, 1–13.

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

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439-453.

Jeffreys, H. (1961). *Theory of probability (3rd edition).* New York: Oxford University Press.

Kass, R. E. (1992). Bayes factors in practice. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *2*, 551–560.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410-423.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187-192.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225-237.

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

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the SavageDickey method. *Cognitive Psychology*, *60*, 158–189.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. (submitted). *Why psychologists must change the way they analyze their data: The case of psi.*

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: Proceedings of the First International Meeting held in Valencia (Spain)* (pp. 585–603). University of Valencia.

## Footnotes

[1] Web applets to compute Bayes factors for paired and grouped $t$-tests may be found at `pcl.missouri.edu/bayesfactor`.

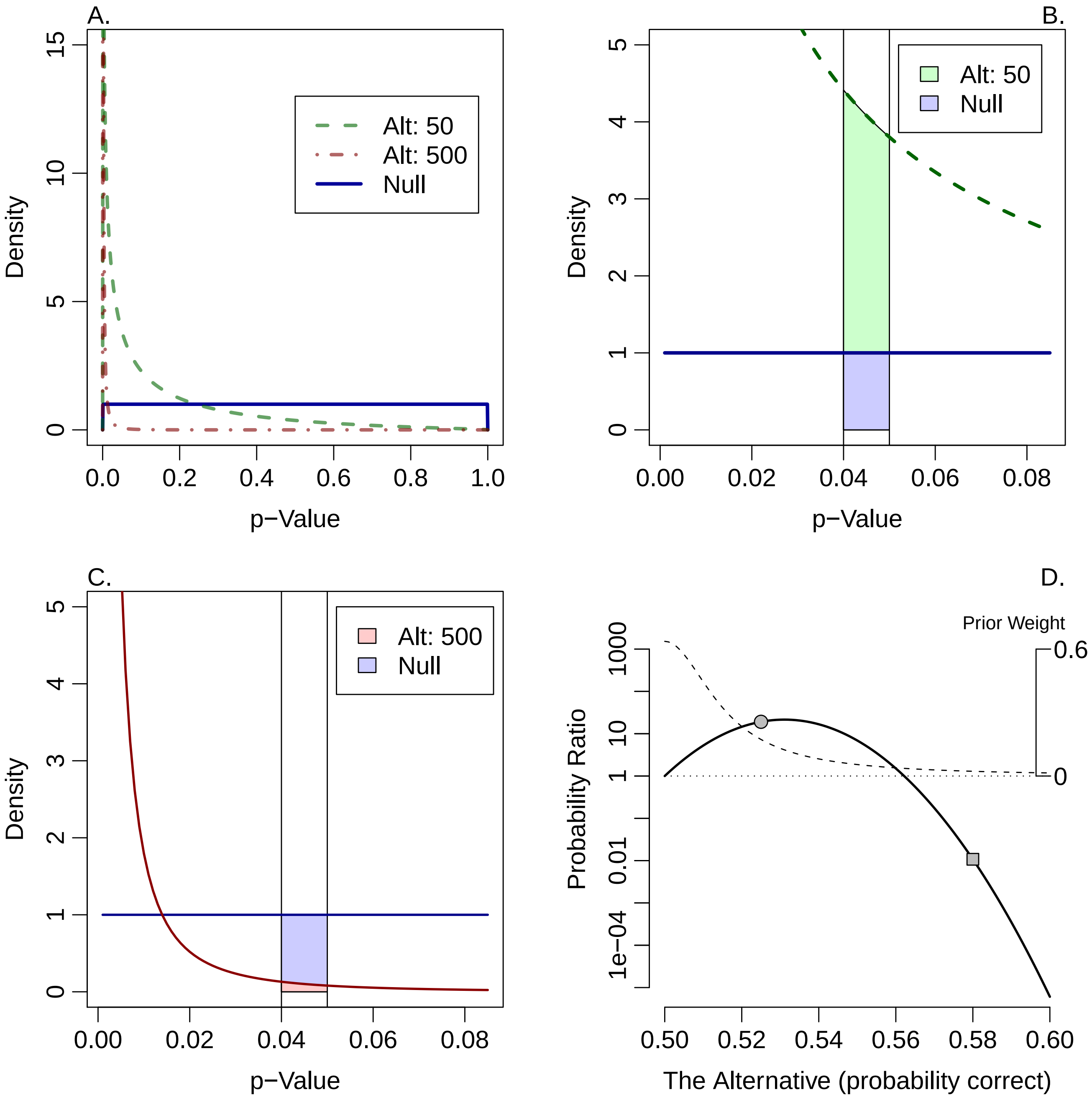
Table 1

*Bayes Factor For Three Feeling-The-Future Hypotheses*

| Stimuli | | Included Experiments | | | | Bayes Factor |
|---|---|---|---|---|---|---|
| Erotic Stimuli | | | | | | 3.23 |
| | Bem's Experiment | #1 | | | | |
| | Sample Size | 100 | | | | |
| | *t*-value | 2.51 | | | | |
| Negative or Positive Stimuli | | | | | | 38.7 |
| | Bem's Experiment | #1 | #2 | #3 | #4 | |
| | Sample Size | 100 | 150 | 97 | 99 | |
| | *t*-value | -.15 | 2.39 | 2.42 | 2.43 | |
| Neutral Stimuli | | | | | | 1.57 |
| | Bem's Experiment | #1 | #8 | #9 | | |
| | Sample Size | 100 | 100 | 50 | | |
| | *t*-value | -.15 | 1.92 | 2.96 | | |

## Figure Captions

*Figure 1.* Significance tests overstate the evidence against the null hypothesis. **A.** The distribution of $p$-values for an alternative with effect-size of .2 (dashed and dashed-dotted lines are for sample sizes of 50 and 500, respectively) and the null (solid lines). **B.** Probability of observing a $p$-value between .04 and .05 for the alternative (effect size = .2) and null for $N = 50$. The probability favors the alternative by a ratio of about 4 to 1. **C.** Probability of observing a $p$-value between .04 and .05 for the alternative (effect size = .2) and null for $N = 500$. The probability favors the null by a factor of 10. **D.** The solid line is the probability of observing a $p$-value of 2.51 for $N = 100$ under the alternative relative to that under the null, as a function of the alternative. The circle and square points highlight the ratios that favor the alternative and null, respectively. The dashed line shows the one-tailed prior distribution used throughout.


A.
Density
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
p−Value
Alt: 50
Alt: 500
Null
B.
Density
0
1
2
3
4
5
0.00
0.02
0.04
0.06
0.08
p−Value
🟩 Alt: 50
🟦 Null
C.
Density
0
1
2
3
4
5
0.00
0.02
0.04
0.06
0.08
p−Value
🟥 Alt: 500
🟦 Null
D.
Prior Weight
0.6
0
Probability Ratio
1000
10
1
0.01
1e−04
0.50
0.52
0.54
0.56
0.58
0.60
The Alternative (probability correct)