Must Psychologists Change the Way They Analyze Their Data?

A Response to Wagenmakers, Wetzels, Borsboom, & Van der Maas (2011)

Daryl J. Bem·

Cornell University

Jessica Utts and Wesley O. Johnson

University of California, Irvine

We agree with Wagenmakers, Wetzels, Borsboom, & van der Maas (2011) that there are advantages to analyzing data with Bayesian statistical procedures, but we argue that they have incorrectly characterized several features of Bem's (2011) psi experiments and have selected an unrealistic Bayesian prior distribution for their analysis, leading them to seriously underestimate the experimental support in favor of the psi hypothesis. We provide an extended Bayesian analysis that displays the effects of different prior distributions on the Bayes factors and conclude that the evidence strongly favors the psi hypothesis over the null. More generally, we believe that psychology would be well served by training future generations of psychologists in the skills necessary to understand Bayesian analyses well enough to perform them on their own data.

In his article "Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect," Bem (2011) performed the standard statistical analyses familiar to most psychologists and concluded that all but one of his nine experiments yielded statistically significant support for the psi hypothesis. Across all nine experiments, the combined (Stouffer) z was 6.66, $p = 1.34 \times 10^{-11}$, with a mean effect size (d) of 0.22.

In their critique, Wagenmakers, Wetzels, Borsboom, & van der Maas (2011) performed a Bayesian analysis on those same data and concluded that "Bem's p values do not indicate evidence in favor of precognition; instead, they indicate that experimental psychologists need to

·Correspondence concerning this article should be addressed to Daryl J. Bem, Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853. d.bem@cornell.edu

change the way they conduct their experiments and analyze their data (abstract, p. xx).” How can we account for this stunning disagreement?

In this response, we seek to answer this question by challenging the particulars of Wagenmakers et al.’s (2011) analysis and to address the more general issue raised by the title of their article, “Why psychologists must change the way they analyze their data.”

Twenty five years ago, Efron (1986) published an article entitled “Why isn’t everyone a Bayesian?” in which he argued that scientists should adopt a combination of Bayesian and the more familiar “frequentist” methods for analyzing data. More recently, Bayesian statisticians Jessica Utts and Wesley Johnson—the second and third authors of this response—have also argued for a Bayesian approach, illustrating their point by performing a Bayesian meta-analysis of 56 experimental studies of telepathy (Utts, Norris, Suess, & Johnson, 2010). (Johnson is also the co-author of the article on Bayesian *t* tests that Wagenmakers et al. [2011] cite as the basis for their own analysis [Gönen, Johnson, Lu, & Westfall, 2005]).

We agree with Wagenmakers et al. (2011) to the extent that we believe psychology would be well served by training future generations of psychologists in the skills necessary to understand Bayesian analyses well enough to perform them on their own data. As Efron (1986) originally warned, however, Bayesian methods require a great deal of thought to apply correctly, and we believe that Wagenmakers et al. (2011) have incorrectly characterized several features of Bem’s experiments, leading them to select a prior distribution that seriously underestimates the support for the psi hypothesis. In particular, we believe they err in (1) treating Bem’s hypotheses as exploratory, thereby requiring two-sided tests, and (2) defining an unreasonable and unreasonably diffuse experimental hypothesis (H_1) to test against the null (H_0) by assuming that we have no prior knowledge on which to base a more reasonable and sharply focused H_1. (An alternative Bayesian analysis by Rouder and Morey [2011] is also critical of Wagenmakers et al.)

The Hypotheses Were Not Exploratory

All nine experiments reported in Bem (2011) tested the single conceptual hypothesis that retroactive or time-reversed versions of common psychological effects would produce the same effects as the standard “forward” versions. Thus, if individuals prefer to approach positive stimuli and avoid negative stimuli, then they should be able to do so precognitively—before the valence of an upcoming stimulus has even been determined (Experiments 1 and 2). If individuals respond more quickly in judging a target stimulus to be pleasant or unpleasant after being exposed to a priming stimulus of the same valence, then they should respond more quickly even if the prime appears after rather than before they judge the target (Experiments 3 and 4).

If repeatedly exposing an individual to a highly arousing stimulus produces habituation so that a strongly negative stimulus becomes less negative and a strongly positive stimulus becomes less positive, then we should observe the same two complementary habituation effects even if the repeated exposures follow rather than precede the assessment of the individual’s affective reactions to the stimuli. (Experiments 5 and 6). And finally, if rehearsing a set of words enhances an individuals’ ability to recall them on a subsequent recall test, then they should display enhanced recall even if the rehearsal takes place after the recall test (Experiments 8 and 9).

These are all unambiguous, highly specific one-sided hypotheses generated from previously established (non-psi) psychological effects. Moreover, four of the nine experiments (Experiments 4, 6, 7, and 9) were themselves replications of the experiments immediately preceding them. Bem’s individual-difference hypothesis that participants high in Stimulus Seeking would show enhanced psi performances in his experiments was conceptually based on Eysenck’s (1966) theorizing about extraversion and psi performance and empirically based on a meta-analysis of 60 experiments in which the correlation between the two could be assessed (Honorton, Ferrari, & Bem, 1992). Across Bem’s experiments, those high in stimulus seeking achieved an effect size (*d*) four times greater than that of other participants (0.43 vs. 0.10), providing additional conceptual coherence to his collected experiments.

Wagenmakers et al. (2011) specifically single out Experiment 1 as exploratory. That experiment was designed to test the hypothesis that participants could identify the future left/right position of an erotic image on the computer screen significantly more frequently than chance. The results showed that they could. The specificity of this hypothesis derives from several earlier "presentiment" experiments (e.g., Radin, 1997) which had demonstrated that participants showed anomalous "precognitive" physiological arousal a few seconds before seeing an erotic image but not before seeing a calm or nonerotic image. Accordingly, Experiment 1 also included randomly interspersed trials with nonerotic images, leaving as an open question whether participants might also be able to anticipate the future left/right positions of these images. They could not, a finding consistent with the results of the presentiment experiments. The important point here is that the central psi hypothesis about erotic images was unambiguous, directional, based on previous research, not conditional on any findings about trials with nonerotic images, and was not formulated from a post hoc exploration of the data. In fact, there was no data exploration that required adjustment for multiple analyses in this or any other experiment.

Perhaps the misunderstanding of this point derives from Bem's practice of applying more than one statistical test to the primary dependent variables (e.g., both a parametric and a nonparametric test) and more than one statistical treatment to some of them (e.g., two different data transformations and two different outlier cutoff criteria on the reaction time data from the priming experiments). But far from being exploratory, these multiple analyses were explicitly confirmatory: They were specifically designed to confirm that the same conclusions held across different statistical treatments of the data.

H_1 Should Reflect Prior Knowledge

Bayesian analyses are designed to pit the null hypothesis (H_0) against a specified experimental hypothesis (H_1). To perform a Bayesian analysis, one must specify two different types of prior belief. The first and most familiar is the prior odds that H_0 is true versus H_1's being true. This is what Wagenmakers et al. (2011, p. xx) set at 99,999,999,999,999,999,999 to 1 odds in favor of H_0. Specifying this type of prior belief gives deniers, believers and everyone in between the opportunity to express a transparent opinion before taking the data into account.

The second part of the prior is more complex and less transparent to those unfamiliar with Bayesian methods. It entails specifying explicitly the effect size probability for both H_0 and H_1. Specifying the effect size for H_0 is easy because it is a single value of 0, but specifying H_1 requires specifying a range of values and a probability distribution over the range for what someone thinks in advance the true effect size might be if H_1 were in fact true.

The "Bayes factor" (BF) indexes the posterior odds of H_1 versus H_0 (or the reverse) after the data are incorporated into the analysis. Numerically it equals the posterior odds for someone whose prior odds were one to one, that is, who initially assigned a prior probability of .5 to both H_0 and H_1. The posterior odds for other prior odds are calculated by simply multiplying those odds by the Bayes factor. Because the Bayes factor is independent of the prior odds, many mistakenly believe that it constitutes an objective assessment of the experimental results, uncontaminated by subjective beliefs. But this is not true because the Bayes factor depends on the specification of H_1.

Accordingly, our second objection to Wagenmakers et al.'s analysis is that their choice of H_1 is unrealistic. Specifically, they assume that we have no prior knowledge of the likely effect sizes that the experiments were designed to detect. As Utts et al. (2010) argue,

> It is rare that we have no information about a situation before we collect data. If we want to estimate the proportion of a community that is infected with HIV, do we really believe it is equally likely to be anything from 0 to 1? If we want to estimate the mean change in blood pressure after 10 weeks of meditation, do we really believe it could be anything from $-\infty$ to $+\infty$? Even the choice of what hypotheses to test, and whether to make them one-sided or two-sided is an illustration of using prior knowledge (p. 2).

In general, we know that effect sizes in psychology typically fall in the range of 0.2 to 0.3. For example, Bornstein's (1989) meta-analysis of 208 mere exposure experiments—the basis of Bem's retroactive habituation experiments—yielded an effect size (*r*) of 0.26. We even have some knowledge about previous psi experiments. The meta-analysis of 56 telepathy studies, cited above, revealed a Cohen's *h* effect size of approximately 0.18 (Utts et al., 2010), and a meta-analysis of 38 "presentiment" studies—from which Bem's experiments 1 and 2 derived—yielded a mean effect size of 0.26 (Mossbridge, Tressoldi, and Utts, 2011).

Surely no reasonable observer would expect effect sizes in laboratory psi experiments to be greater than 0.8—what Cohen (1988) calls a large effect. (Cohen notes that even a medium effect of 0.5 "is large enough to be visible to the naked eye" [p. 26].) Yet the "default prior" that Wagenmakers et al. (2011) use (known as the standard Cauchy distribution) has probability 0.57 that the absolute value of the effect size exceeds 0.8. It even places probability of 0.12 on effect sizes with absolute values exceeding 5.0, and probability of 0.06 on effect sizes with absolute values exceeding 10! If the effect sizes were really that large, there would be no debate about the reality of psi. Thus, the prior distribution they have placed on the possible effect sizes under H_1 is wildly unrealistic.

When the null hypothesis is sharply defined but the prior distribution on the alternative hypothesis is diffused over a wide range of values, it is more likely that the probability of *any* observed data will be higher under the null hypothesis than under the alternative. This is known as the Lindley-Jeffreys paradox: A frequentist analysis that yields strong evidence in support of the alternative hypothesis can be contradicted by an inappropriate Bayesian analysis that concludes that the same data are more likely under the null. Christensen et al. (2011) discuss an example comparable to the analysis by Wagenmakers et al. (2011), noting that "the moral of the Lindley-Jeffreys paradox is that if you pick a stupid prior, you can get a stupid posterior (p. 60)."

Testing Alternative Prior Distributions for H_1

We now examine what happens when more realistic prior distributions are used to define H_1. Because the tests reported in Bem's original article were justifiably one-sided, we begin by using prior distributions that include positive effect sizes only. Modifying Gönen et al. (2005) for a one-sided situation, we used a half-normal distribution starting at 0 for our alternative priors. The only parameter required for this distribution is the spread, and so we specified the value corresponding to the 90th percentile of the distribution.

We call our first alternative prior the "Knowledge-Based prior," because it reflects what we already know about effect sizes typically observed in psychological research, including previous psi research. Using the earlier outcomes for guidance, we set the 90th percentile to be an effect size of 0.5. Someone with this prior believes that if precognition is real, the true effect size is less than or equal to 0.5 with probability 0.9. For comparison, we also display the results of using a two-sided version of the Knowledge-Based prior, which reflects the alternative hypothesis that the *absolute value* of the true effect size is less than or equal to 0.5 with probability 0.9. For the "Skeptic's Prior" we set the 90th percentile to be only 0.05, reasoning that a skeptic might think an effect between 0 and 0.05 might occur due to possible artifacts or even a very small but unimportant psi effect, but that an effect size greater than 0.05 would be unlikely. And finally, we calculate the half-Cauchy prior, which is equivalent to the one-sided version or upper half of the two-sided prior used by Wagenmakers et al. (2011).

As in Rouder et al. (2009) and consistent with the way Bem generated his data, we assume that the data values are normal with mean μ and variance σ^2, with the Jeffreys' prior (1961) serving as the standard reference prior for the variance in this model. For all our computations we used Markov chain Monte Carlo simulations with the statistical software WinBUGS (Lunn, Thomas, Best & Spiegelhalter, 2000) to get numerical approximations.

First, we computed the Bayes factor for H_1 to H_0 for each of the nine experiments under the four priors. (Wagenmakers et al. [2011] actually present Bayes factors of H_0 to H_1, but it is easier here to interpret the reciprocal, H_1 to H_0. See, for example, Bayarri and Berger's [1991] Bayesian analysis of psi data.) Using the assumption that the effect sizes under H_1 for the separate experiments are independent and are drawn from a single effect size distribution, we also calculated a Bayes factor for the nine experiments combined by computing the product of the separate Bayes factors under each of the priors. And finally, for these Bayes factors we calculated the associated posterior probability that H_0 is true for all of the experiments (when the prior probability on all H_0 being simultaneously true is .5). In this analysis, we assume that either all null hypotheses are true or all alternative hypotheses are true. The results are shown in Table 1. Wagenmakers et al.'s (2011) results are shown in the last column, and the combined Bayes factors and posterior probabilities on H_0 are shown in the bottom two rows.

The most striking finding is that under the knowledge-based prior, the Bayesian analysis yields exactly the same conclusions as Bem's (2011) original frequentist analysis. Using Jeffreys' (1961) verbal labels for characterizing the size of a Bayes factor (BF), every experiment (except Experiment 7, as in Bem [2011]) shows either "strong" (BF > 10) or "substantial" (BF > 3) evidence in favor of H_1. The combined Bayes factor of 5,184,907 easily exceeds his criterion for "extreme" evidence in favor of H_1 (BF > 100), with a posterior probability on the composite H_0 of 1.9×10^{-7}. Even the two-sided version of that same prior H_1 distribution yields "extreme" evidence in favor of the psi alternative, with a posterior probability on the composite H_0 of 7.3 × 10^{-5}. Finally, both the Skeptic's prior and the one-sided Cauchy prior also yield "extreme" evidence in favor H_1. Only the two-sided Cauchy prior used by Wagenmakers et al. (2011) fails to show strong support for the psi hypothesis.

Table 1

Bayes Factors (BF) H_1 to H_0 for Five Prior Distributions on H_1

Experiment	**Knowledge-Based Prior**	**Knowledge-Based Prior (Two-Sided)**	**Skeptic's Prior**	**Cauchy One-Sided**	**Cauchy Two-sided (Wagenmakers et al.)**[a]
1	9.62	4.94	1.86	3.09	1.64
2	6.89	3.45	2.07	2.04	1.05
3	10.42	5.35	1.86	3.43	1.82
4	3.41	1.76	1.59	1.03	0.58
5	5.37	2.74	1.72	1.70	0.88
6	7.40	3.78	2.11	2.21	1.10[b]
7	0.90	0.50	1.45	0.23	0.13
8	3.16	1.62	1.57	0.92	0.47
9	19.48	10.12	1.64	9.85	5.88
Combined	**5,184,907**	**13,668.9**	**154.3**	**174.4**	**0.632**
Posterior *pr* all H_0	**1.9 × 10^{-7}**	**7.3 ×10^{-5}**	**0.0064**	**0.0057**	**0.61**

[a] Wagenmakers et al. (2011) reported Bayes factors of H_0 to H_1, so the figures in this column are the reciprocals (H_1 to H_0) of their numbers.

[b] Wagenmakers et al. evaluated two separate *t* tests reported by Bem for Experiment 6; we used the combined *t* test and have updated their Bayes factor to correspond to that combined *t* test.

In an online appendix to their article, Wagenmakers et al. (2011) claim to show that their conclusions are robust across different priors for H_1, but they restrict their discussion to two-sided Cauchy priors that are still very diffuse, and they never consider the combined evidence across all of the experiments. For example, if they had simply considered a two-sided Cauchy prior that places 90% of the probability on effect sizes with absolute value less than 0.5—like our two-sided Knowledge-Based prior distribution—they, too, would have discovered "extreme" evidence in favor of H_1, namely, a composite Bayes factor of 1,964 and a posterior probability on the composite H_0 of 0.0005.

Critics of using Bayesian analyses for psi hypotheses frequently point out the reductio ad absurdum case of the extreme skeptic who believes psi to be impossible, that is, who holds the prior probability of 0 for the psi alternative. In this case, no finite amount of data can raise the posterior probability in favor of the alternative hypothesis above 0 or, alternatively, lower the posterior probability in favor of the null below 1. This extreme case does, however, raise the question of how close to 0 the prior probability for the alternative would need to be to maintain a posterior probability close to 0.95 for the null. For the Knowledge-based prior, one's prior probability that the alternative is true would have to be 10^{-8} (or $1 - 10^{-8}$ that the null is true). Thus, when taking the combined data into account it would take a mighty strong prior belief in the null hypothesis to retain even a reasonably high posterior belief in it. Of course Wagenmakers et al. (2011) admit that they do indeed have more than a mighty strong belief in the null hypothesis ($1 - 10^{-20}$), so even the posterior probability of 1.9×10^{-7} obtained with the Knowledge-based prior would not convince them, as it might convince a skeptic with a less extreme position.

Here Be Dragons

In choosing to present the standard frequentist analysis of his data, Bem (2011) noted that

> There are, of course, more sophisticated statistical techniques available…but they do not yet appear to be widely familiar to psychologists and are not yet included in popular statistical computer packages, such as SPSS. I have deliberately not used them for this article. It has been my experience that the use of complex or unfamiliar statistical procedures in the reporting of psi data has the perverse effect of weakening rather than strengthening the typical reader's confidence in the findings.… [T]his is understandable. If one holds low Bayesian a priori probabilities about the existence of psi—as most academic psychologists do—it might actually be more logical from a Bayesian perspective to believe that some unknown flaw or artifact is hiding in the weeds of…an unfamiliar statistical analysis than to believe that genuine psi has been demonstrated (p. xx).

Ironically, Wagenmaker et al.'s (2011) critique itself provides an illuminating example of how hidden assumptions can lurk "in the weeds" of an unfamiliar statistical analysis—albeit here in the service of proclaiming the null hypothesis.

Medieval maps used to mark unknown or unexplored territories with the warning "Here Be Dragons." Until a new generation of psychologists becomes as familiar with Bayesian analyses as their mentors have become with frequentist analyses, a similar warning would seem appropriate.

References

Bayarri, M. J. & Berger, J. (1991). Comment *Statistical Science, 6,* 379-382.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407-425.

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

Christensen, R., Johnson, W., Branscum, A. & Hanson, T. E. (2011). *Bayesian ideas and data analysis: An introduction for scientists and statisticians*, Boca Raton, FL: Chapman & Hall.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician, 40*, 1-5.

Eysenck, H. J. (1966). Personality and extra-sensory perception. *Journal of the Society for Psychical Research, 44*, 55–71.

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two–sample *t* test. *The American Statistician*, *59*, 252–257.

Honorton, C., Ferrari, D. C., & Bem, D. J. (1992). Extraversion and ESP performance: Meta-analysis and a new confirmation. In L. A. Henkel & G. R. Schmeidler (Eds.), *Research in parapsychology 1990* (pp. 35–38). Metuchen, NJ: Scarecrow Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press, Clarendon Press.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS —a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10,* 325–337.

Mossbridge, J, Tressoldi, P, and Utts, J. (2011). Physiological anticipation of unpredictable stimuli: A meta-analysis. Unpublished manuscript.

Radin, D. I. (1997). Unconscious perception of future emotions: An experiment in presentiment. *Journal of Scientific Exploration, 11*, 163–180.

Rouder, J. N. & Morey, R. D. (2011). *A Bayes-factor meta analysis of Bem's ESP claim*. Manuscript submitted for publication.

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

Utts, J., Norris, M., Suess, E, & Johnson, W. (2010). The strength of evidence versus the power of belief: Are we all Bayesians? In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society*. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute. http://www.stat.auckland.ac.nz/~iase/publications.php[© 2010 ISI/IASE]

Wagenmakers, EJ., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, *100*, 426-432.