Journal of Personality and Social Psychology (in press) © 2011 American Psychological Association
http://www.apa.org/pubs/journals/psp/index.aspx 0022-3514/11/$12.00 DOI: 10.1037/a0024777

REPLY

Must Psychologists Change the Way They Analyze Their Data?·

Daryl J. Bem
Cornell University

Jessica Utts and Wesley O. Johnson
University of California, Irvine

Wagenmakers, Wetzels, Borsboom, & van der Maas (2011) argue that psychologists should replace the familiar "frequentist" statistical analyses of their data with Bayesian analyses. To illustrate their argument, they reanalyzed a set of psi (ESP) experiments published recently in this journal by Bem (2011), maintaining that, contrary to his conclusion, his data do not yield evidence in favor of the psi hypothesis. We argue that they have incorrectly selected an unrealistic prior distribution for their analysis and that a Bayesian analysis using a more reasonable distribution yields strong evidence in favor of the psi hypothesis. More generally, we argue that there are advantages to Bayesian analyses that merit their increased use in the future. However, as Wagenmakers et al.'s analysis inadvertently revealed, they contain hidden traps that need to be better understood before being more widely substituted for the familiar frequentist analyses currently employed by most research psychologists.

Keywords: Bayesian analysis, statistics, psi, ESP

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. Recently, Wagenmakers, Wetzels, Borsboom, & van der Maas (2011) argued that psychologists must actually replace those familiar methods with Bayesian analyses. To illustrate their argument, they reanalyzed a set of psi (ESP) experiments published in this journal by Bem (2011) and argued that, contrary to his conclusion, "Bem's *p* values do not indicate evidence in favor of precognition" (p. 426). In this brief response, we examine their analysis.

· Daryl J. Bem, Department of Psychology, Cornell University; Jessica Utts and Wesley O. Johnson, Department of Statistics, University of California, Irvine.

We thank statistician Ronald Christensen for his helpful comments during the preparation of this article.

Correspondence concerning this article should be addressed to Daryl J. Bem, Department of Psychology, Uris Hall, Cornell University, Ithaca, NY 14853. E-mail: d.bem@cornell.edu

The term *psi* denotes anomalous processes of information transfer that are currently unexplained in terms of known physical or biological mechanisms. One variant of psi is the anomalous retroactive influence of some future event on an individual's current responses. In his article, Bem (2011) reported nine experiments testing the hypothesis that retroactive or time-reversed versions of four common psychological effects would produce the same effects as the standard "forward" versions. For example, it is known that rehearsing a set of verbal materials enhances an individuals' ability to recall them on a subsequent free recall test. In a time-reversed experiment and its replication, Bem demonstrated that individuals will display enhanced recall even if the rehearsal takes place after the recall test has been administered. (In criticizing Bem's research as exploratory, Wagenmakers et al., 2011, failed to acknowledge that four of his nine experiments were actually replications of others in the set.)

In reporting his results, 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, $p = 2.68 \times 10^{-11}$, two-tailed, with a mean effect size (*d*) of 0.22.

We are not opposed to Bayesian analyses. In fact, Jessica Utts and Wesley Johnson—the second and third authors of this response—are Bayesian statisticians who have themselves analyzed psi data. For example, they and two co-authors performed a Bayesian meta-analysis of 56 experimental studies of telepathy (Utts, Norris, Suess, & Johnson, 2010). Utts and two co-authors performed a Bayesian meta-analysis of 38 "presentiment" studies—from which two of Bem's experiments derived (Mossbridge, Tressoldi, and Utts, 2011). And finally, Johnson is the co-author of the article on Bayesian *t* tests that Wagenmakers et al. (2011) cited as the basis for their own analysis (Gönen, Johnson, Lu, & Westfall, 2005).

As Efron (1986) originally warned, however, it requires careful thought to apply Bayesian methods correctly, and we believe that Wagenmakers et al. (2011) have not done so. (Rouder and

Morey, 2011, who also performed a Bayesian analysis of Bem's 2011 data, are also critical of the Wagenmakers et al. analysis.)

The Challenge of Specifying the Experimental Hypothesis

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 rather than H_1 is true. It is here that Wagenmakers et al. (2011) formally expressed their prior skepticism about the existence of psi by setting these odds at 99,999,999,999,999,999,999 to 1 in favor of H_0. Specifying this type of prior belief gives deniers, believers, and everyone in between the opportunity to express an explicit opinion before taking the data into account.

The second prior belief that must be specified is more complicated and not widely known to those unfamiliar with the details of Bayesian analysis. This is the explicit specification of a probability distribution of effect sizes across a range for both H_0 and H_1. Specifying the effect size for H_0 is simple because it is a single value of 0, but specifying H_1 requires specifying a probability distribution across a range of what the effect size might be if H_1 were in fact true.

Another element of Bayesian analysis not widely familiar to research psychologists is the "Bayes Factor," a number that 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 itself is independent of the prior odds, it can easily be mistaken for an objective assessment of the experimental results, uncontaminated by subjective beliefs. But this is not true because the Bayes factor depends on the prior specification of H_1.

Accordingly, our critique of Wagenmakers et al.'s (2011) analysis is that their choice of H_1 is unrealistic. In particular, they assumed that we have no prior knowledge of the likely effect sizes that the experiments were explicitly designed to detect. As Utts et al. (2010) argued,

> 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. A survey of "one hundred years of social psychology" that catalogued 25,000 studies of eight million people yielded a mean effect size (*r*) of .21 (Richard, Bond, & Stokes-Zoota, 2003). An example relevant to Bem's (2011) retroactive habituation experiments is Bornstein's (1989) meta-analysis of 208 mere exposure studies, which yielded an effect size (*r*) of 0.26.

We even have some knowledge about previous psi experiments. The Bayesian meta-analysis of 56 telepathy studies, cited above, revealed a Cohen's *h* effect size of approximately 0.18 (Utts et al., 2010), and the meta-analysis of 38 "presentiment" studies, also cited above, yielded a mean effect size of 0.28 (Mossbridge, et al., 2011).

Consequently, no reasonable observer would ever expect effect sizes in laboratory psi experiments to be greater than 0.8—what Cohen (1988) terms 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 prior distribution for H_1 that Wagenmakers et al. (2011) adopted places a probability of .57 on effect sizes that equal or exceed 0.8. It even places a probability of .06 on effect sizes exceeding 10. If effect sizes were really that large, there would be no debate about the reality of psi. Thus, the prior distribution Wagenmakers et al. placed on the possible effect sizes under H_1 is wildly unrealistic.

Their unfortunate choice has major consequences for their conclusions about Bem's data. Whenever the null hypothesis is sharply defined but the prior distribution on the alternative hypothesis is diffused over a wide range of values, as it is in the distribution adopted by Wagenmakers et al. (2011), it boosts the probability that *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 experimental hypothesis can be contradicted by a misguided Bayesian analysis that concludes that the same data are more likely under the null. Christensen , Johnson, Branscum, and Hanson (2011) discussed an analysis comparable to that of Wagenmakers et al., 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 a Knowledge-Based Distribution for H_1

We now examine what happens when a more realistic prior distribution is used to define H_1. We call our distribution the "knowledge-based" prior because it reflects what we already know about effect sizes typically observed in psychological research, including previous psi research. We selected a normal distribution centered on 0 for our alternative prior, and the only parameter required to specify this distribution is the spread. Using the earlier outcomes for guidance, we set the 90th percentile of the absolute values to be an effect size of 0.5. That is, someone with this prior believes that if psi is real, the probability is .9 that the absolute value of the true effect size will be less than or equal to 0.5.[1]

Next, we computed the Bayes factor for H_1 to H_0 for each of Bem's (2011) nine experiments under this prior. (Wagenmakers et al., 2011, actually presented Bayes factors of H_0

[1] Following Rouder, Speckman, Dongchu, Morey, and Iverson (2009), 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.

to H_1, but it is easier here to interpret the reciprocal, H_1 to H_0. See, for example, the Bayesian analysis of psi data in Bayarri and Berger, 1991.) 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. 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. Bem's frequentist results are shown in the first data column and Wagenmakers et al.'s results are shown in the last column. The combined Bayes factors and posterior probabilities on H_0 for all nine experiments are shown in the bottom two rows.

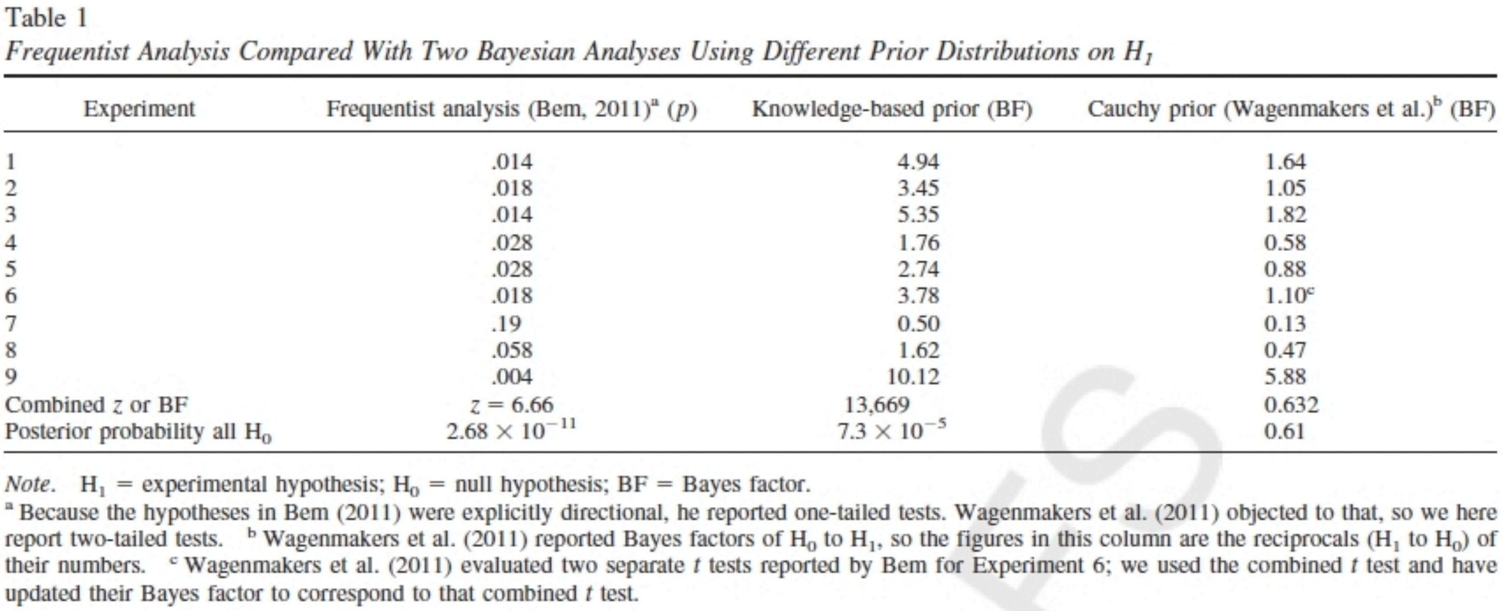

Table 1
Frequentist Analysis Compared With Two Bayesian Analyses Using Different Prior Distributions on H_1

Experiment	Frequentist analysis (Bem, 2011)[a] (p)	Knowledge-based prior (BF)	Cauchy prior (Wagenmakers et al.)[b] (BF)
1	.014	4.94	1.64
2	.018	3.45	1.05
3	.014	5.35	1.82
4	.028	1.76	0.58
5	.028	2.74	0.88
6	.018	3.78	1.10[c]
7	.19	0.50	0.13
8	.058	1.62	0.47
9	.004	10.12	5.88
Combined z or BF	$z = 6.66$	13,669	0.632
Posterior probability all H_0	2.68×10^{-11}	7.3×10^{-5}	0.61

Note. H_1 = experimental hypothesis; H_0 = null hypothesis; BF = Bayes factor.
[a] Because the hypotheses in Bem (2011) were explicitly directional, he reported one-tailed tests. Wagenmakers et al. (2011) objected to that, so we here report two-tailed tests. [b] 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. [c] Wagenmakers et al. (2011) 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.

The main finding in Table 1 is that under the knowledge-based prior, the Bayesian analysis yields the same overall conclusion as Bem's (2011) original frequentist analysis. In fact, if we adopt Wagenmakers et al.'s (2011) own verbal labels for characterizing the size of a Bayes factor (BF)—which they, in turn, adapted from Jeffreys (1961)—five of the nine experiments individually yield either “strong” (BF > 10) or “substantial” (BF > 3) evidence in favor of H_1. Most important, the combined Bayes factor of 13,669 greatly exceeds their own criterion for “extreme” evidence in favor of H_1 (BF > 100), with a posterior probability on the composite H_0 of 7.3×10^{-5}. Only the diffuse prior used by Wagenmakers et al. (2011)—known as the standard Cauchy distribution—fails to show strong support for the psi hypothesis.

In an online appendix to their article, Wagenmakers et al. (2011) claimed to show that their conclusions are robust across different priors for H_1, but they continued to confine their consideration to diffuse Cauchy priors. If they had simply considered a Cauchy prior analogous to our knowledge-based prior (i.e., one that places a 90% probability on effect sizes with absolute value less than 0.5) 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 frequently point out the reductio ad absurdum case of the extreme skeptic who declares psi (or any testable phenomenon) to be impossible, that is, who holds the prior probability of 0 for the psi alternative. In this case, the Bayesian formula implies that no finite amount of data can raise the posterior probability in favor of the psi hypothesis above 0 or, alternatively, lower the posterior probability in favor of the null below 1. The critics point out that this effectively confers analytic legitimacy on the most anti-scientific stance.

More realistically, all an extreme skeptic needs to do is to set his or her prior odds on the psi alternative sufficiently low so as to rule out the probative force of any data that could reasonably be proffered. This raises the question in the present case of how close to 0 the prior probability for the psi alternative would need to be to maintain a posterior probability in favor of the null close to .95. For the knowledge-based prior with a one-sided alternative in favor of psi, one’s prior probability that the psi alternative is true would have to be 10^{-8} or lower.

Thus, when taking the combined data into account, it would take very strong initial skepticism regarding psi to retain a reasonably high posterior belief in the null. Of course Wagenmakers et al. (2011) admitted that their a priori belief in the psi alternative is, indeed, very close to zero (10^{-20}), so even the posterior probability of 7.3×10^{-5} for the null obtained with the knowledge-based prior fails to exceed their threshold for being convinced.

Must Psychologists Change the Way They Analyze Their Data?

We now return to the original question raised by Wagenmakers et al. (2011): Must psychologists change the way they analyze their data? We believe that scientific questions addressed with modeled data can almost always be approached with either frequentist or Bayesian methods and that this question is similar to asking whether traditional chalk-and-talk lecturers must change the way they lecture. The answer to both is "no" if they are content with their current practices and prefer not to adopt new, more versatile tools." Precisely because Bayesian methods expand the modeling options, there are now thousands of Bayesian analyses published in the scientific literature and an explosion of Bayesian methodology in most data-oriented disciplines.

An anonymous reviewer of this article commented that

> I have great sympathy for the Bayesian position.…The problem in implementing Bayesian statistics for scientific publications, however, is that such analyses are inherently subjective, by definition…with no objectively right answer as to what priors are appropriate. I do not see that as useful scientifically. …[I]t is unclear to me how we can have agreed upon priors for a collective such as the body of psychological researchers.

We believe that this comment reflects a misunderstanding about Bayesian statistics that may be widespread among psychologists. Eminent statistician George Box has noted that "essentially all models are wrong, but some are useful" (Wasserstein, 2010). In fact all model-building in statistics is inherently subjective, and we believe there is great utility in specifying an explicit prior distribution in a statistical analysis.

There are three kinds of prior distributions that can be adopted in a Bayesian analysis, each with a different goal. Some Bayesian analyses use “reference” priors that are designed to minimize the effect of the prior on the conclusions. Others use “objective” priors, designed to result in inferences that satisfy frequentist criteria. Finally, there are “subjective” priors, which are warranted when there is relevant prior information that can be quantified. This is what we have adopted here in our reanalysis of Bem’s (2011) data, incorporating prior information consistent with previous psychological literature on effect sizes. (We also performed sensitivity analyses in which we perturbed these priors modestly to confirm that the conclusions do not change.) Readers who disagree with our prior are free to disagree with our conclusions. In contrast, Wagenmakers et al (2011) have specified a prior distribution that is not based on any prior knowledge. Instead, it is a prior that produces the Lindley-Jeffreys paradox because it posits unrealistically large departures from the null under the psi alternative. As we note above, more than half of their posited distribution lies beyond a gigantic effect size of 0.8. Having failed to see such implausible departures, they concluded that the data are more consistent with the null than the alternative. Readers are now in a position to decide for themselves which prior seems the more plausible.

This debate is an excellent illustration of how science works. Different individuals working on the same scientific problem come to different conclusions based on their own assumptions and models—which Bayesian methods make explicit. Such disagreements persist until there is sufficient information available to convince the broader scientific community where the truth lies. Many will prefer the comfort zone of *p* values, which have played a valuable role in statistical analyses for many decades. But the statistical world is changing, and it seems likely that Bayesian methods will be playing an increasing role in the analysis of all types of data, including psychological data.

In reporting his data, Bem (2011) presented a more pragmatic argument for choosing the standard frequentist analysis:

> 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…. 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.… This is understandable. If one holds low Bayesian [prior] 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. 420).

Ironically, Wagenmaker et al.'s (2011) critique itself provides an illuminating example of how hidden flaws or artifacts can lurk "in the weeds" of an unfamiliar statistical analysis—albeit here in the service of defending 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 the hidden traps of Bayesian analyses as their mentors have become with those of frequentist analyses, a similar warning would seem appropriate.

References

Bayarri, M. J. & Berger, J. (1991). Comment. *Statistical Science, 6*, 379-382. doi:10.1214/ss/1177011578

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. doi:10.1037/a0021524

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin, 106*, 265–289. doi:10.1037/0033-2909.106.2.265

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. doi:10.2307/2683105

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two–sample t test. *The American Statistician, 59*, 252–257. doi:10.1198/000313005X55233

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. doi:10.1023/A:1008929526011

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

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331-363. doi:10.1037/1089-2680.7.4.331

Rouder, J. N. & Morey, R. D. (2011). A Bayes-factor meta analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*. Advance online publication. DOI: 10.3758/s13423-011-0088-7

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. doi:10.3758/PBR.16.2.225

Utts, J., Norris, M., Suess, E, & Johnson, W. (2010, July). *The strength of evidence versus the power of belief: Are we all Bayesians?* Paper presented at the Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia. Retrieved from http://www.stat.auckland.ac.nz ~iase/publications.php

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. doi:10.1037/a0022790

Wasserstein, R. (2010). George Box: A model statistician. *Significance, 7(3),* 134-135. doi:10.1111/j.1740-9713.2010.00442.x

Received February 5, 2011
Revision received May 5, 2011
Accepted May 9, 2011