

# Yes, Psychologists Must Change the Way They Analyze Their Data: Clarifications for Bem, Utts, and Johnson (2011)

Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, Rogier Kievit, & Han L. J. van der Maas  
University of Amsterdam

## Abstract

Does psi exist? In a widely publicized article featuring nine experiments with over one thousand participants, Bem (in press) claimed that future events retroactively affect people's responses. In a response, we pointed out that Bem's analyses were partly exploratory. Moreover, we reanalyzed Bem's data using a default Bayesian  $t$ -test and showed that Bem's evidence for psi is weak to nonexistent. A robustness analysis confirmed our skeptical conclusions. Recently, Bem, Utts, and Johnson (2011) question several aspects of our analysis. In this brief reply we clarify our analysis procedure and demonstrate that our arguments still hold.

**Keywords:** Confirmatory Experiments, Bayesian Hypothesis Test, ESP.

## The History and the Hype

In a recent article for *Journal of Personality and Social Psychology*, Bem (in press) presented nine experiments that test for the presence of psi. Specifically, Bem's experiments were designed to assess the hypothesis that future events affect people's thinking and people's behavior in the past (henceforth precognition). Bem argued that in eight out of the nine experiments, the data supported the presence of precognition, that is, one-sided  $p$  values were smaller than .05.

Bem's findings—and, perhaps more importantly, the fact that they were going to be published in a major journal—created a storm of media attention. In the *New York Times*, several researchers voiced strong opinions: Dr. Ray Hyman, a long-time critic of ESP research, questioned the quality of the refereeing process as he believed that the publication of Dr. Bem's article was “(...) pure craziness (...) an embarrassment for the

---

This version was last updated with minor changes on February 18th, 2011. This research was supported by Vidi grants from the Dutch Organization for Scientific Research (NWO). Correspondence concerning this article may be addressed to Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands. Email address: ej.wagenmakers@gmail.com.

entire field”<sup>1</sup>, and Dr. Douglas Hofstadter argued for “(...) a cutoff for craziness, and when that threshold is exceeded, then the criteria for publication should get far, far more stringent.” Bem’s article was also discussed in *Science* (Miller, 2011) and many other media throughout the world. A Google search on “Bem” and “feeling the future” generates over 50,000 hits.<sup>2</sup> Bem himself appeared on the popular US television show *The Colbert Report*, where the host described Bem’s work as “extrasensory pornception” referring to the fact that Experiment 1 in Bem (in press) found that precognition was present only for erotic pictures. In the *New York Times*, Bem was quoted as saying “What I showed was that unselected subjects could sense the erotic photos, but my guess is that if you use more talented people, who are better at this, they could find any of the photos.”

Some months before Bem’s research started to attract a lot of media attention we wrote a response that criticized Bem’s work on several counts. This response was submitted to JPSP and published in the same issue (i.e., Wagenmakers, Wetzels, Borsboom, & van der Maas, in press). In this response, we first noted that the analysis of the experiments had been partly exploratory, whereas the statistical analysis assumed a fully confirmatory approach. That is, we argued that Bem had used the data twice: once to discover an interesting result, and then to test it. In support of our claim, we pointed to several instances where it was clear that the analysis had been exploratory.

Next we used Bayes theorem to argue that the bar for publishing should be set higher for claims that are outlandish or improbable. Third, we used a default Bayesian  $t$  test (Rouder, Speckman, Sun, Morey, & Iverson, 2009) to highlight that the one-sided  $p$  values used by Bem overestimate the evidence against the null; in fact, our default test indicated little evidence in favor of precognition—only one of Bem’s nine experiments yielded data substantially more likely under  $H_1$  (i.e., the hypothesis of precognition) than under  $H_0$ .

It is important to note that our default Bayesian test does not depend at all on the prior probability that one may assign  $H_1$ . Therefore, it is certainly not true that our Bayesian analysis simply confirms our initial bias against precognition, as some bloggers mistakenly believed. Instead, the result of our Bayesian test is known as the *Bayes factor*, and with respect to prior assumptions it only depends on the effect size  $\delta$  expected under  $H_1$  (see also Liang, Paulo, Molina, Clyde, & Berger, 2008). In what follows, we will denote the prior distribution for effect size under  $H_1$  as  $p(\delta | H_1)$ .

The default assumption we made about  $p(\delta | H_1)$  was based on a long tradition in Bayesian statistics where prior distributions are constructed from general desiderata (Jeffreys, 1961). The advantage that this brings is that the Bayesian analysis is fully objective (Berger, 2004) and avoids subjective specification of the expected effect sizes under  $H_1$ . We realized that the default choice leads to a conservative test. Indeed, our abstract stated that “(...) in order to convince a skeptical audience of a controversial claim, one needs to conduct strictly confirmatory studies and analyze the results with statistical tests that are conservative rather than liberal.”

Despite the advantages of an objective test, we also realized that the choice of  $p(\delta | H_1)$  could be disputed. We therefore carried out a robustness analysis in which we systematically

---

<sup>1</sup>Dr. Hyman did not question the publication of a parapsychological article as such. Instead, Dr. Hyman was puzzled that JPSP had accepted an article with so many departures from accepted methodological practice (Dr. Hyman, personal communication).

<sup>2</sup>Query issued on 15 February 2011.

varied the scale parameter for  $p(\delta | H_1)$ , and reported the results in an online appendix.<sup>3</sup> These results showed that for a wide range of different, non-default prior distributions on effect size the evidence for precognition is either non-existent or negligible.

The penultimate section of our response provided guidelines on confirmatory research. We stressed how important it is that research on precognition is conducted in the context of an adversarial collaboration, that is, a collaboration with a qualified skeptic (e.g., Diaconis, 1991).

Throughout our response, we argued that our critique was not meant to attack research on psi. The last paragraph of our response is particularly clear on the broader consequences of the debate:

“It is easy to blame Bem for presenting results that were obtained in part by exploration; it is also easy to blame Bem for possibly overestimating the evidence in favor of  $H_1$  because he used  $p$  values instead of a test that considers  $H_0$  vis-a-vis  $H_1$ . However, Bem played by the implicit rules that guide academic publishing—in fact, Bem presented many more studies than would usually be required. It would therefore be mistaken to interpret our assessment of the Bem experiments as an attack on research of unlikely phenomena; instead, our assessment suggests that something is deeply wrong with the way experimental psychologists design their studies and report their statistical results. It is a disturbing thought that many experimental findings, proudly and confidently reported in the literature as real, might in fact be based on statistical tests that are explorative and biased (...). We hope the Bem article will become a signpost for change, a writing on the wall: psychologists must change the way they analyze their data.”

The broader impact of our response to Bem has been described as “the Bayesian bomb”.<sup>4</sup> Consistent with this assessment, Wetzels et al. (in press) presented default Bayes factors for all 855  $t$  tests reported in the 2007 volumes of *Psychonomic Bulletin & Review* and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. The results showed that for 70% of the data sets for which  $p$  values range from .01 to .05, the Bayes factor indicated that the evidence in favor of  $H_1$  is “anecdotal” in the sense that the data are less than three times more likely under  $H_1$  than under  $H_0$ .

#### The Complaints by Bem, Utts, and Johnson (2011)

A recent rebuttal by Bem et al. (2011)<sup>5</sup> questions several aspects of our response outlined above. We disagree with several of their points, but we also believe that something good may come out of this debate, at least for the field of psi.

Below we discuss the Bem et al. (2011) rebuttal in terms of four central complaints. The first is that Bem (in press) did *not* explore the data when he analyzed his results. We argue that this general statement fails to address our detailed points of critique, that in earlier work Bem himself argued strongly in favor of exploration, and that the Bem

<sup>3</sup>Available on the first author’s website or at [http://www.ruudwetzels.com/articles/Wagenmakersetal\\_robust.pdf](http://www.ruudwetzels.com/articles/Wagenmakersetal_robust.pdf).

<sup>4</sup>George van Hal, NWT Magazine.

<sup>5</sup>Downloaded from <http://dbem.ws/ResponsetoWagenmakers.pdf> on February 15th, 2011.

experiments show a strong negative correlation between sample size and effect size (as first pointed out by Dr. Hyman, personal communication).

The second complaint is that in Bem's experiments a one-sided test is more appropriate than a two-sided test. Although we generally agree that a Bayesian one-sided test can be entirely appropriate (e.g., Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009) the danger of a one-sided test is that it can be abused in the absence of strong *a priori* expectations to create an overly optimistic impression of the true evidence in favor of the hypothesis under consideration. We will illustrate this danger with three experiments reported in Bem (in press).

The third complaint is that our default prior distribution on effect size,  $p(\delta | H_1)$ , was too wide and assigned too much weight to implausibly high values of effect size. As indicated above, we had already addressed this issue in our robustness analysis. However, we do appreciate the proposal for a specific prior distribution that can now be used to compute subjective or informed Bayes factors in the field of psi. Perhaps future studies will use this prior to evaluate the evidence in favor or against precognition and psi. We examine a two-sided version of the proposed prior distribution in detail in the penultimate section of this paper.

The fourth complaint is that evidence should be combined across studies. We agree that, in an ideal world, combining information across multiple studies is useful. However, this is not a perfect world, and as stated in our response:

(...)we have assessed the evidential impact of Bem's experiments in isolation. It is certainly possible to combine the information across experiments, for instance by means of a meta-analysis (Storm, Tressoldi, & Di Risio, 2010; Utts, 1991). We are ambivalent about the merits of meta-analyses in the context of psi: one may obtain a significant result by combining the data from many experiments, but this may simply reflect the fact that some proportion of these experiments suffer from experimenter bias and excess exploration. When examining different answers to criticism against research on psi, Price (1955, p. 367) concluded "But the only answer that will impress me is an adequate experiment. Not 1000 experiments with 10 million trials and by 100 separate investigators giving total odds against chance of  $10^{1000}$  to 1—but just one good experiment."

We also note that Bem's article would most likely not have been published if it had to back away from the claim that the experiments showed *independent* evidence for precognition, i.e., when considered in isolation. JPSP does not publish many experiments with 200 participants that yield inconclusive results.

We now deal with each of the complaints in detail. The reader who is bored can safely skip to the Conclusion section.

#### *Complaint 1: There Really Was No Exploration*

Bem et al. (2011) deny that there was any exploration in the Bem (in press) experiments. They argue that the hypotheses were all based on prior research, and that even though multiple analyses were conducted, these analyses served to confirm the same point. This statement contrasts sharply with reality.

First of all, Bem et al. (2011) do not address the specific points of concern that we raised in four paragraphs of our response. For example, it is completely unclear why gender effects were tested in the first place, as Bem (in press) explicitly states that “the psi literature does not reveal any systematic sex differences in psi ability”. In addition, our experience is that psychologists explore their data at least to some extent. When Bem et al. (2011) claim not to have explored the data at all, they effectively state that the research by Bem (in press) is the pinnacle of confirmatory research. This impression is inconsistent with a painfully detailed analysis of the Bem experiments by James Alcock.<sup>6</sup> Moreover, this impression is also inconsistent with the quotation from the Bem chapters on writing that we presented in our response:

“The conventional view of the research process is that we first derive a set of hypotheses from a theory, design and conduct a study to test these hypotheses, analyze the data to see if they were confirmed or disconfirmed, and then chronicle this sequence of events in the journal article. (...) But this is not how our enterprise actually proceeds. Psychology is more exciting than that (...)” (Bem, 2000, p. 4).

Unfortunately, Bem et al. (2011) chose not to elaborate on the extent to which the philosophy behind this quotation (and others) discredits the conclusions from all statistical analysis, Bayesian, frequentist, or otherwise.

As a final indication that the results from Bem (in press) were obtained from exploration, Ray Hyman (personal communication) noted that in the Bem study the low effect sizes tended to occur in experiments with many participants. Figure 1 shows this association (see also Hyman, 1985). How can we explain this if the experiments were purely confirmatory?

In sum, Bem et al. (2011) fail to address the questions about exploration that we raised in our response. In addition, the Bem experiments with many participants show smaller effects than those with fewer participants. This strongly suggests that exploration (perhaps through optional stopping) did take place.

### *Complaint 2: A One-Sided Test is More Appropriate Than a Two-Sided Test*

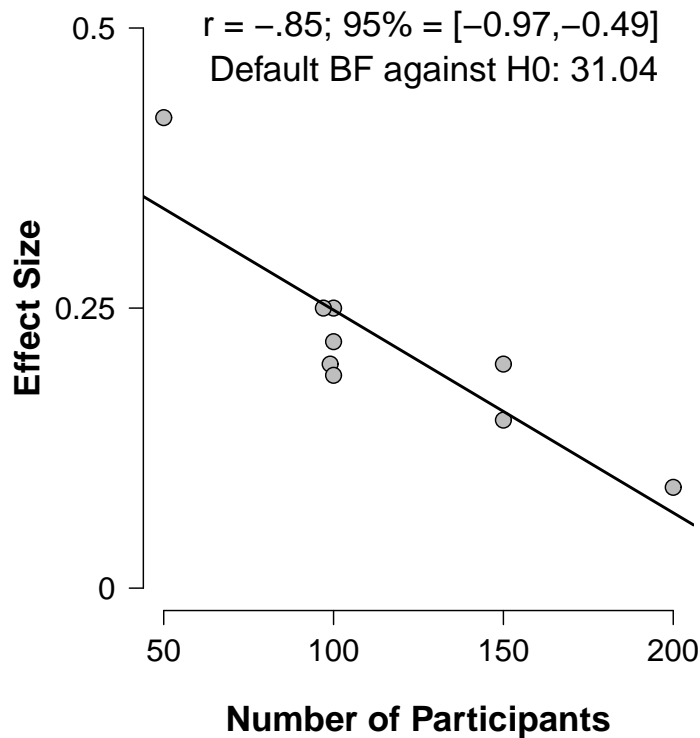
Bem et al. (2011) argue that the tests for precognition in the Bem studies should be one-sided, not two-sided. As pointed out above, the main problem with one-sided tests is that they may be used to bias the results. That is, a researcher without strong a priori expectations may await the data and select the one-sided test that produces the most convincing result. In fact, this disadvantage is illustrated in the very paper that Bem et al. (2011) seek to defend.

The problem concerns Experiments 5, 6, and 7 and it is perhaps best illustrated with a comment from Rouder and Morey (2011)<sup>7</sup>, who also advocated the use of a one-sided test but excluded these experiments from consideration:

---

<sup>6</sup>Available at [http://www.csicop.org/specialarticles/show/back\\_from\\_the\\_future](http://www.csicop.org/specialarticles/show/back_from_the_future). Bem’s response and Alcock’s reply can also be found online.

<sup>7</sup>Downloaded from <http://pcl.missouri.edu/sites/default/files/rouder-morey.pdf> on February 15th, 2011.



*Figure 1.* In the Bem experiments, low effect sizes are associated with high number of participants (Hyman, personal communication; see also Hyman, 1985). The default Bayes factor is due to Jeffreys, 1961, pp. 289-292.

“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.”

In the presence of such ambiguity, we feel it is prudent to simply stick to a two-sided test. After all, the case for precognition would be quite weak if depended on the use of a one-sided test instead of a two-sided test.

*Complaint 3: The Default Priors are Implausible*

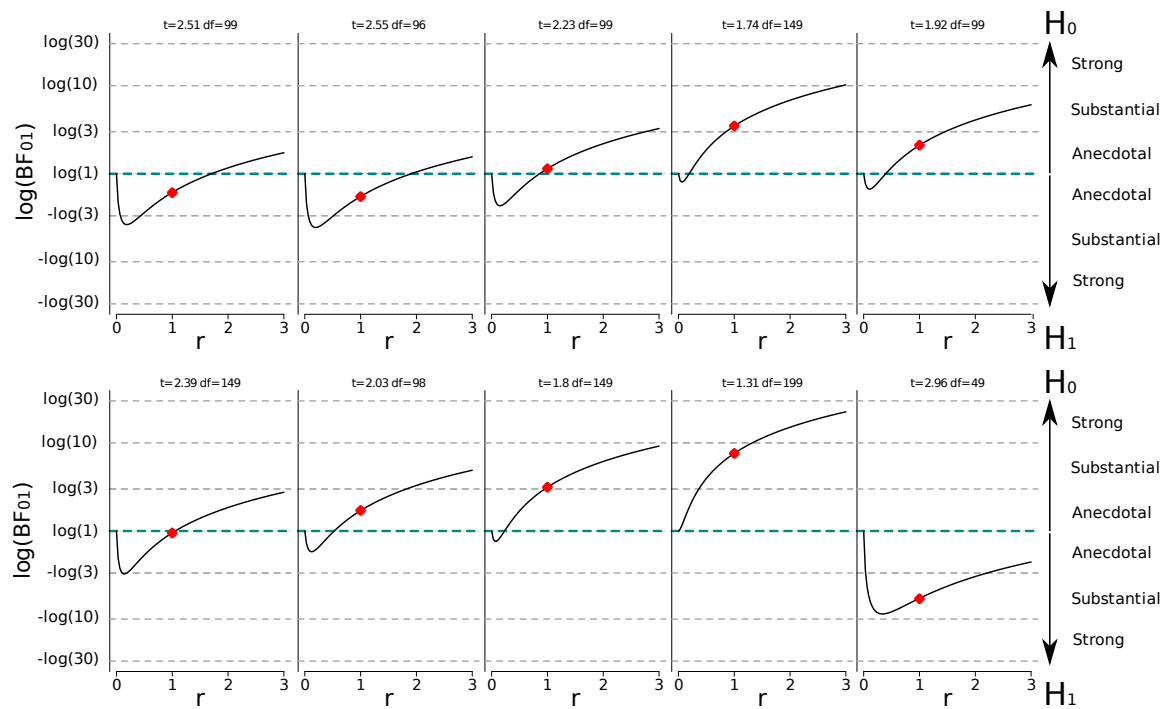
Bem et al. (2011) believe that our default prior for effect size  $\delta$  is unrealistic, because it attaches too much weight to values of  $\delta$  that are relatively large. As mentioned above, we feel that an important advantage of the default test is that it is objective and relatively conservative. We would also like to point out that we did not invent this default prior ourselves; it is in fact a standard choice for variable selection in Bayesian regression analysis (Liang et al., 2008), and Rouder et al. (2009) have promoted its use for the  $t$  test. However, we realized that researchers might want to specify different priors, and this is why we conducted a robustness analysis in which we manipulated the scale parameter  $r$  of the prior distribution. Low values of  $r$  indicate that the effect sizes are expected to be close to zero.

Our robustness analysis showed that for a wide range of different, non-default prior distributions on effect size  $\delta$  the evidence for precognition is either non-existent or negligible. Our results are replotted in Figure 2 and, as we explained in our online appendix:

“Note that Figure 2 plots the Bayes factor such that the scale of evidence in favor of  $H_0$  is visually equivalent to the scale of evidence in favor of  $H_1$ . Also note that when  $r = 0$ ,  $H_0 = H_1$ , and the Bayes factor indicates that the evidence is perfectly ambiguous (i.e.,  $BF_{01} = 1$ ). The different panels in Figure 2 indicate that our choice for the default prior does not affect our conclusions. In fact, the red dot—the result of our default test—seems to provide a relatively accurate summary of the evidence. Yes, it is true that for very small values of  $r$  the evidence is occasionally in favor of  $H_1$ , but—and this is the crucial point—only for the bottom right panel is the evidence clearly in favor of  $H_1$ . That is, in the bottom right panel the maximum Bayes factor is almost 1/10, meaning that the observed data are about 10 times more likely under  $H_1$  than they are under  $H_0$ , given of course that the prior scale parameter  $r$  is chosen a posteriori, something that greatly biases the Bayes factor in favor of  $H_1$ . For 7 out of the remaining 9 other panels, even the maximum Bayes factor indicates only “anecdotal” evidence (i.e., evidence worth “no more than a bare mention”, that is, the data are less than 3 times more likely under  $H_1$  than under  $H_0$ ). This leaves the top-left two panels, for which the maximum Bayes factor does reach the criterion for “substantial” evidence; however, it does so only just, and only for very specific values of the scale parameter. Again, the default test (indicated by the red dot) seems to provide a reasonable indication of the evidence.

In sum, we conclude that our results are robust to different specifications of the scale parameter for the effect size prior under  $H_1$ . This reinforces our general argument that  $p$ -values may strongly overstate the evidence against  $H_1$ .”

Finally, one could easily disagree with the way in which Bem et al. (2011) determined their “Knowledge-Based” prior for effect size  $\delta$  (henceforth the BUJ prior). For instance, Bem et al. (2011) point out that most effect sizes in psychology are between 0.2 and 0.3. Instead, Figure 3 shows that this is incorrect, and that many published effect sizes are greater than 1. Also, Bem et al. (2011) inform the estimates of effect size for psi by reviews of experiments on telepathy and presentiment. Skeptical researchers may not consider these reviews to be valid indicators of the effect size for psi at all. Instead, skeptical researchers



*Figure 2.* A robustness analysis for the data from Bem (in press). The Bayes factor  $BF_{01}$  is plotted as a function of the scale parameter  $r$  of the Cauchy prior for effect size under  $H_1$ . The red dot indicates the result from the default prior, the horizontal green line indicates complete ambiguous evidence, and the horizontal grey lines demarcate the different qualitative categories of evidence (see Wagenmakers et al., in press). Importantly, the results in favor of  $H_1$  are never compelling, except perhaps for the bottom right panel.

may well believe that the reviews of telepathy and presentiment have only summarized the extent to which experiments were manipulated, the data explored, and the statistics rigged.

Despite these reservations about the way the BUJ prior was motivated, we are willing to except that in the case of psi one can make a legitimate case for a subjective prior that is more narrowly peaked around zero than the default Cauchy. We are, however, wary of using the one-sided version of the BUJ prior, for reasons outlined in the previous section. In the penultimate section of this paper, we present a robustness analysis of the two-sided BUJ prior and show that their choice of its variance is very close to the choice that maximizes the evidence against  $H_0$ ; in addition, we show that even the unrealistic maximum evidence is still unimpressive on an experiment-by-experiment basis.

Of course, the dilemma regarding priors can be avoided by choosing the default option. This option leads to a test that may be conservative—however, if one finds an effect despite the conservative nature of the test, one can be relatively certain that something is going on. The data of Bem et al. (2011) are simply not convincing enough to pass this test.

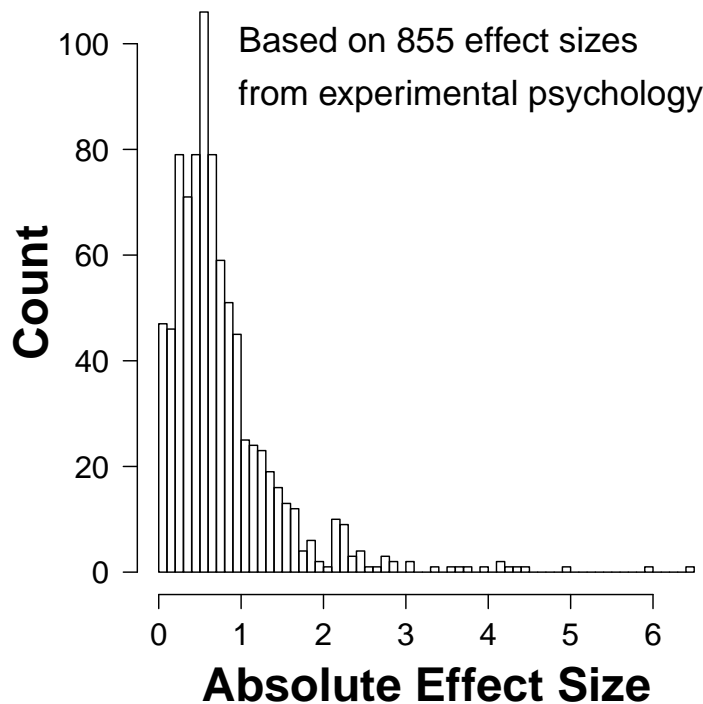


Figure 3. Histogram for 855 effect sizes computed from all  $t$  tests reported in the 2007 volume of *Psychonomic Bulletin & Review* and the *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Data from Wetzels et al., in press.

#### *Complaint 4: Evidence Should be Combined Across Studies*

The final complaint of Bem et al. (2011)—one that truly makes a difference—is that the evidence should be combined across experiments by multiplying the Bayes factors. We believe this is a major concession. Apparently unsatisfied with the evidence that each study provides on its own, Bem et al. (2011) resort to the last refuge of the refuted: consider all experiments simultaneously.

It seems to us that the experiments originally reported by Bem (in press) were meant to stand on their own, and for good reason. If Bem (in press) in fact thought that the results were *not* independent sources of evidence for psi, why then were they reported as such? If the results in fact hinge on a meta-analysis, then why did Bem (in press) report piecemeal analyses? Regardless, we take notice of meta-analyses based on careful, confirmatory experiments that were carried out in the presence of qualified skeptics; a post-hoc meta-analysis of experiments that were originally presented as independent evidence does not convince us (see also Hyman, 2010, p. 489).

Consider then the outcome of the Bayes factor analysis with the one-sided knowledge-based BUJ prior. Based on the arguments by Rouder and Morey (2011), we eliminate from consideration Experiments 5, 6, and 7. Next, the Bayes factors in favor of  $H_1$  for Experiments 4 and 8 are 3.41 and 3.16, respectively. This means that the data are about 3 times more likely under  $H_1$  than they are under  $H_0$ . This is not very impressive and these numbers contrast with the significant  $p$  values reported in Bem (in press). A two-sided test would firmly place the evidence in the category “anecdotal”. This then leaves Experiments 1, 2, and 9. Anybody can make up their own mind and decide whether three experiments with Bayes factors of 9.62, 6.89, and 19.48 are enough to seriously entertain the suggestion that psi exists. We can only speak for ourselves and say that even if these experiments had been confirmatory, this amount of evidence falls far short of what is required.

On a technical note, we are also puzzled by the fact that Bem et al. (2011) choose to multiply the Bayes factors from the individual experiments.<sup>8</sup> It seems more informative to do a hierarchical analysis and treat the experiments as a random effect. However, given the doubtful origin of the data and the experimental design we suggest that additional analysis efforts are probably misplaced.

### A Robustness Analysis of the BUJ Prior

As noted above, skeptical researchers may believe the construction of the BUJ prior is somewhat ad-hoc. It is particularly worrisome, these researchers may argue, that the prior was constructed after the data had been already collected, analyzed, and debated. To study the effect of prior specification in more detail, we again conducted a robustness analysis, as we did before, but now for the Gaussian distribution. In contrast to Rouder and Morey (2011), we wanted to analyze all experiments and therefore used a two-sided test. Thus, we assume  $\delta \sim N(0, \sigma)$  and we consider the different values for the Bayes factor as a function of  $\sigma$ .

The result is shown in Figure 4. It is clear that even the maximum Bayes factor in favor of  $H_1$  fails to make a big impression, except perhaps for Experiment 9. It is also clear that the two-sided BUJ prior, indicated by the red dot, is remarkably close to the maximum Bayes factor for many of the experiments. We certainly do not mean to suggest that Bem et al. (2011) cherry-picked their prior to give a convincing result, but we do feel that in cases such as this a simple robustness analysis allows for a more comprehensive assessment of the evidence.

The green dot corresponds to  $\sigma = .95$ . This prior is constructed from the empirical data shown in Figure 3 using the same procedure used by Bem et al. (2011). That is, Bem et al. (2011) obtained a  $\sigma$  (i.e., the red dot) by assuming that 90% of effects are smaller than 0.5; we obtained our  $\sigma$  (i.e., the green dot) by assuming that 40% of effects are smaller than 0.5 – this 40% was based on 855  $t$  tests from experimental psychology reported in Wetzels et al. (in press). We note that the conclusions from our data-based prior are very similar to the ones drawn from our default prior.

Of course, our data-based prior can be disputed too; for instance, one may argue that the data from Wetzels et al. (in press) probably show a lot of underreporting (e.g.,

<sup>8</sup>In their own response to Bem (in press), Rouder and Morey (2011) state “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.”

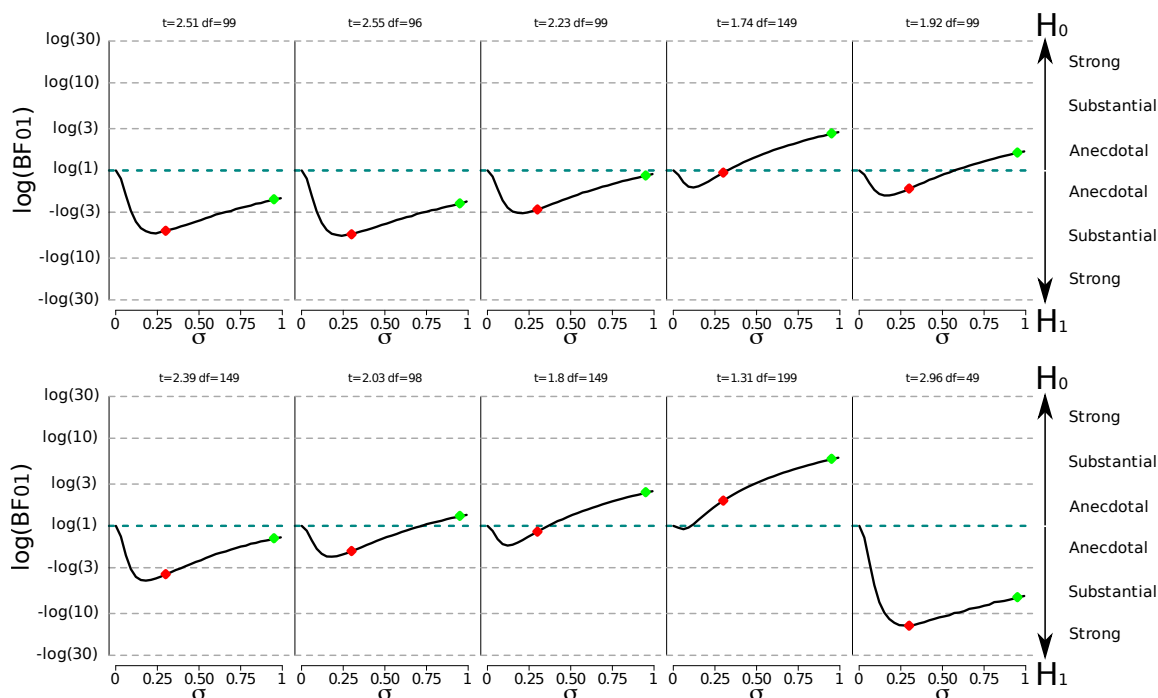


Figure 4. Bayes factors for the two-sided Normal prior as a function of standard deviation  $\sigma$ . The red dot is the BUJ prior, remarkably close to the maximum possible Bayes factor against  $H_0$ . The green dot is the prior constructed from the 855 tests analyzed by Wetzels et al. (in press), using the same procedure as described in Bem et al. (2011).

some articles state “all other effects were not significant”, which leaves open how many other effects were tested and what effect sizes were found). And indeed, it may be better to estimate  $\sigma$  and the amount of underreporting at the same time. This more elaborate modeling of the effect size distribution is ongoing in our lab, but we felt it was important to demonstrate that the available data appear to be inconsistent with the broad claim by Bem et al. (2011) that most effect sizes in psychology are between 0.2 and 0.3.

## Conclusion

The complaints listed by Bem et al. (2011) are either overstated or wrong. We have presented additional evidence that the data from Bem (in press) have been obtained through exploration, and noted that our original worries about exploration have simply not been addressed. As noted by Rouder and Morey (2011), the plea for a one-sided test is inconsistent with Bem’s Experiments 5, 6, and 7 and suggests further exploration has taken part. The proposal to multiply the Bayes factors across all experiments is a measure of desperation, and an implicit acknowledgement that the data from each experiment separately do not convince. From our perspective, the only fair proposal in Bem et al. (2011) is to consider an informed prior on effect size. A cautious researcher might still want to use the two-sided version of this BUJ prior. Also, objective default Bayes factors that fail to support  $H_1$  will

not convince many researchers to put a lot of trust in the results.

In some unintended ways, the alternative analysis proposed by Bem et al. (2011) has strengthened our case. Even with their informed prior, the results from the two-sided version will not differ much from the conclusion that we already drew. Given the fact that we had previously reported a robustness analysis this should not come as a surprise. Also, the individual experiment analysis by Bem et al. (2011) highlights the tension between  $p$  values and Bayes factors, where the former overestimate the evidence against the null. This tension does not arise because the priors are extremely uninformative, as Bem et al. (2011) suggest. Instead, this tension is an inalienable aspect of Bayes factors in general—it always occurs when  $n$  increases but the  $p$ -value is held constant (e.g., Lindley, 1957; Wagenmakers & Grünwald, 2006).<sup>9</sup>

At the end of their rebuttal, Bem et al. (2011) mention dragons, referring to the dangers of using Bayesian analyses without the proper background knowledge. We cannot take this remark seriously. Not only did we use a default Bayesian test, but we also conducted an elaborate robustness analysis to consider the generality of our conclusions. The real dragons in academia are not those who seek to improve the way we draw conclusions from data; instead, they are those who do not want to understand the difference between exploratory and confirmatory analyses, potentially wasting other people's time and money on replication attempts that are doomed to fail. We conclude that the rebuttal by Bem et al. (2011) underscores what the (Bem, in press) already showed: psychologists must change the way they analyze their data.

---

<sup>9</sup>Perhaps Bem et al. (2011) meant to refer to the paradox that the Bayes factor can favor the null hypothesis irrespective of the data—this occurs when the prior for the parameter of interest is extremely uninformative.

## References

- Bem, D. J. (2000). Writing an empirical article. In R. J. Sternberg (Ed.), *Guide to publishing in psychology journals* (pp. 3–16). Cambridge: Cambridge University Press.
- Bem, D. J. (in press). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*.
- Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? A response to wagenmakers, wetzels, borsboom, & van der Maas (2011). Manuscript submitted for publication.
- Berger, J. (2004). The case for objective Bayesian analysis. *Bayesian Analysis*, 1, 1–17.
- Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, 106, 265–289.
- Diaconis, P. (1991). Comment. *Statistical Science*, 6, 386.
- Hyman, R. (1985). The ganzfeld psi experiment: A critical appraisal. *Journal of Parapsychology*, 49, 3–49.
- Hyman, R. (2010). Meta-analysis that conceals more than it reveals: Comment on Storm et al. (2010). *Psychological Bulletin*, 136, 486–490.
- Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.
- 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.
- Miller, G. (2011). News of the week: ESP paper rekindles discussion about statistics. *Science*, 331, 272–273.
- Price, G. R. (1955). Science and the supernatural. *Science*, 122, 359–367.
- Rouder, J. N., & Morey, R. D. (2011). An assessment of the evidence for feeling the future with a discussion of Bayes factor and significance testing. Manuscript submitted for publication.
- 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 & Review*, 16, 225–237.
- Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin*, 136, 471–485.
- Utts, J. (1991). Replication and meta-analysis in parapsychology (with discussion). *Statistical Science*, 6, 363–403.
- Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing. *Psychological Science*, 17, 641–642.
- Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, 60, 158–189.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (in press). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*.

- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (in press). Statistical evidence in experimental psychology: An empirical comparison using 855  $t$  tests. *Perspectives on Psychological Science*.
- Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian  $t$  test. *Psychonomic Bulletin & Review*, *16*, 752–760.