Journal of Personality and Social Psychology

# Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011)

Eric–Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han L. J. van der Maas

Online First Publication, January 31, 2011. doi: 10.1037/a0022790

Journal of Personality and Social Psychology


0022-3514/11/$12.00 DOI: 10.1037/a0022790

# Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011)

Eric–Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han L. J. van der Maas
University of Amsterdam

Does psi exist? D. J. Bem (2011) conducted 9 studies with over 1,000 participants in an attempt to demonstrate that future events retroactively affect people's responses. Here we discuss several limitations of Bem's experiments on psi; in particular, we show that the data analysis was partly exploratory and that one-sided *p* values may overstate the statistical evidence against the null hypothesis. We reanalyze Bem's data with a default Bayesian *t* test and show that the evidence for psi is weak to nonexistent. We argue 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. We conclude that Bem's *p* values do not indicate evidence in favor of precognition; instead, they indicate that experimental psychologists need to change the way they conduct their experiments and analyze their data.

*Keywords:* confirmatory experiments, Bayesian hypothesis test, ESP

Bem (2011) presented nine experiments that test for the presence of psi.[1] The experiments were designed to assess the hypothesis that future events affect people's thinking and people's behavior in the past (henceforth, precognition). As indicated by Bem, precognition—if it exists—is an anomalous phenomenon, because it conflicts with what we know to be true about the world (e.g., weather forecasting agencies do not employ clairvoyants, casinos make profits). In addition, psi has no clear grounding in known biological or physical mechanisms.[2]

Despite the lack of a plausible mechanistic account of precognition, Bem (2011) was able to reject the null hypothesis of no precognition in eight out of nine experiments. For instance, in Bem's first experiment 100 participants had to guess the future position of pictures on a computer screen, left or right. And indeed, for erotic pictures, the 53.1% mean hit rate was significantly higher than chance, $t(99) = 2.51$, $p = .01$.

Bem (2011) took these findings to support the hypothesis that people "use psi information implicitly and nonconsciously to enhance their performance in a wide variety of everyday tasks" (p. 16). In further support of psi, Utts (1991, p. 363) concluded in a *Statistical Science* review article that "the overall evidence indicates that there is an anomalous effect in need of an explanation" (but see Diaconis, 1978; Hyman, 2007). Do these results mean that psi can now be considered real, replicable, and reliable?

We think that the answer to this question is negative and that the take-home message of Bem's (2011) research is in fact of a completely different nature. One of the discussants of the Utts review paper made the insightful remark that "parapsychology is worth serious study . . . If it is wrong [i.e., psi does not exist], it offers a truly alarming massive case study of how statistics can mislead and be misused" (Diaconis, 1991, p. 386). And this, we suggest, is precisely what Bem's research really shows. Instead of revising our beliefs regarding psi, Bem's research should instead cause us to revise our beliefs on methodology: The field of psychology currently uses methodological and statistical strategies that are too weak, too malleable, and offer far too many opportunities for researchers to befuddle themselves and their peers.

The most important flaws in the Bem (2011) experiments, discussed below in detail, are the following: (a) confusion between exploratory and confirmatory studies; (b) insufficient attention to the fact that the probability of the data given the hypothesis does not equal the probability of the hypothesis given the data (i.e., the fallacy of the transposed conditional); (c) application of a test that overstates the evidence against the null hypothesis, an unfortunate tendency that is exacerbated as the number of participants grows large. Indeed, when we apply a Bayesian *t* test (Gönen, Johnson, Lu, & Westfall, 2005; Rouder, Speckman, Sun, Morey, & Iverson,

Eric–Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han L. J. van der Maas, Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands.

This research was supported by Vidi grants from the Dutch Organization for Scientific Research. We thank Rogier Kievit and Jan de Ruiter for constructive discussions.

Correspondence concerning this article should be addressed to Eric–Jan Wagenmakers, University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands. E-mail: ej.wagenmakers@gmail.com

[1] The preprint on which this article was based was downloaded September 25, 2010, from http://dbem.ws/FeelingFuture.pdf

[2] Some argue that modern theories of physics are consistent with precognition. We cannot independently verify this claim but note that work on precognition is seldom published in reputable physics journals (in fact, we failed to find a single such publication). But even if the claim were correct, the fact that an assertion is consistent with modern physics does not make it true. The assertion that the CIA bombed the twin towers is consistent with modern physics, but this fact alone does not make the assertion true. What is needed in the case of precognition is a plausible account of the process that leads future events to have perceptual effects in the past.

2009) to quantify the evidence that Bem presented in favor of psi, the evidence is sometimes slightly in favor of the null hypothesis and sometimes slightly in favor of the alternative hypothesis. In almost all cases, the evidence falls in the category "anecdotal," also known as "worth no more than a bare mention" (Jeffreys, 1961).

We realize that the above flaws are not unique to the experiments reported by Bem (2011). Indeed, many studies in experimental psychology suffer from the same mistakes. However, this state of affairs does not exonerate the Bem experiments. Instead, these experiments highlight the relative ease with which an inventive researcher can produce significant results even when the null hypothesis is true. This evidently poses a significant problem for the field and impedes progress on phenomena that are replicable and important.

## Problem 1: Exploration Instead of Confirmation

In his well-known book chapters on writing an empirical journal article, Bem (2000, 2003) rightly called attention to the fact that psychologists do not often engage in purely confirmatory studies. That is,

> 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)

How is it then that psychologists analyze their data? Bem noted that senior psychologists often leave the data collection to their students and made the following recommendation:

> To compensate for this remoteness from our participants, let us at least become intimately familiar with the record of their behavior: the data. Examine them from every angle. Analyze the sexes separately. Make up new composite indexes. If a datum suggests a new hypothesis, try to find further evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don't like, or trials, observers, or interviewers who gave you anomalous results, place them aside temporarily and see if any coherent patterns emerge. Go on a fishing expedition for something—anything—interesting. (Bem, 2000, pp. 4–5)

We agree with Bem (2000) in the sense that empirical research can benefit greatly from a careful exploration of the data; dry adherence to confirmatory studies stymies creativity and the development of new ideas. As such, there is nothing wrong with fishing expeditions. But it is vital to indicate clearly and unambiguously which results are obtained by fishing expeditions and which results are obtained by conventional confirmatory procedures. In particular, when results from fishing expeditions are analyzed and presented as if they had been obtained in a confirmatory fashion, the researcher is hiding the fact that the same data were used twice: first to discover a new hypothesis and then to test that hypothesis. If the researcher fails to state that the data have been so used, this practice is at odds with the basic ideas that underlie scientific methodology (for a detailed discussion, see Kerr, 1998).

Instead of presenting exploratory findings as confirmatory, one should ideally use a two-step procedure. First, in the absence of strong theory, one can explore the data until one discovers an interesting new hypothesis. But this phase of exploration and discovery needs to be followed by a second phase, one in which the new hypothesis is tested against new data in a confirmatory fashion. This is particularly important if one wants to convince a skeptical audience of a controversial claim: After all, confirmatory studies are much more compelling than exploratory studies. Hence, explorative elements in the research program should be explicitly mentioned, and statistical results should be adjusted accordingly. In practice, this means that statistical tests should be corrected to be more conservative.

The Bem (2011) experiments were at least partly exploratory. For instance, Bem's Experiment 1 tested not just erotic pictures but also neutral pictures, negative pictures, positive pictures, and pictures that were romantic but nonerotic. Only the erotic pictures showed any evidence for precognition. But now suppose that the data would have turned out differently and instead of the erotic pictures, the positive pictures would have been the only ones to result in performance higher than chance. Or suppose the negative pictures would have resulted in performance lower than chance. It is possible that a new and different story would then have been constructed around these other results (Bem, 2003; Kerr, 1998). This means that Bem's Experiment 1 was to some extent a fishing expedition, an expedition that should have been explicitly reported and should have resulted in a correction of the reported $p$ value.

Another example of exploration comes from Bem's (2011) Experiment 3, in which response time (RT) data were transformed using either an inverse transformation (i.e., 1/RT) or a logarithmic transformation. These transformations are probably not necessary, because the statistical analysis were conducted on the level of participant mean RT; one then wonders what the results were for the untransformed RTs, results that were not reported.

Furthermore, in Bem's (2011) Experiment 5, the analysis shows that "women achieved a significant hit rate on the negative pictures, 53.6%, $t(62) = 2.25$, $p = .014$, $d = 0.28$; but men did not, 52.4%, $t(36) = 0.89$, $p = .19$, $d = 0.15$" (p. 10). But why test for gender in the first place? There appears to be no good reason. Indeed, Bem himself stated that "the psi literature does not reveal any systematic sex differences in psi ability" (p. 10).

Bem's (2011) Experiment 6 offers more evidence for exploration, as this experiment again tested for gender differences but also for the number of exposures: "The hit rate on control trials was at chance for exposure frequencies of 4, 6, and 8. On sessions with 10 exposures, however, it fell to 46.8%, $t(39) = -2.12$, two-tailed $p = .04$" (p. 11). Again, conducting multiple tests requires a correction.

These explorative elements are clear from Bem's (2011) discussion of the empirical data. The problem runs deeper, however, because we simply do not know how many other factors were taken into consideration only to come up short. We can never know how many other hypotheses were in fact tested and discarded; some indication is given above and in Bem's section The File Drawer. At any rate, the foregoing suggests that strict confirmatory experiments were not conducted. This means that the reported $p$ values are incorrect and need to be adjusted upward.

## Problem 2: Fallacy of the Transposed Conditional

The interpretation of statistical significance tests is liable to a misconception known as the fallacy of the transposed conditional. In this fallacy, the probability of the data given a hypothesis (e.g., $p(D|H)$, such as the probability of someone being dead given that he was lynched, a probability that is close to 1) is confused with the probability of the hypothesis given the data (e.g., $P(H|D)$, such as the probability that someone was lynched given that he is dead, a probability that is close to zero).

This distinction provides the mathematical basis for Laplace's principle that extraordinary claims require extraordinary evidence. This principle holds that even compelling data may not make a rational agent believe that psi exists (see also Price, 1955). Thus, the prior probability attached to a given hypothesis affects the strength of evidence required to make a rational agent change his or her mind.

Suppose, for instance, that in the case of psi we have the following hypotheses:

$$H_0 = \text{Precognition does not exist;}$$

$$H_1 = \text{Precognition does exist.}$$

Our personal prior belief in precognition is very low; two reasons for this are outlined below. We accept that each of these reasons can be disputed by those who believe in psi, but this is not the point—we do not mean to disprove psi on logical grounds. Instead, our goal is to indicate why most researchers currently believe psi phenomena are unlikely to exist.[3]

As a first reason, consider that Bem (2011) acknowledges that there is no mechanistic theory of precognition (for a discussion, see Price, 1955). This means, for instance, that we have no clue about how precognition could arise in the brain. Neither animals nor humans appear to have organs or neurons dedicated to precognition, and it is unclear what electrical or biochemical processes would make precognition possible. Note that precognition conveys a considerable evolutionary advantage (Bem, 2011), and one might therefore assume that natural selection would have led to a world filled with powerful psychics (i.e., people or animals with precognition, clairvoyance, psychokinesis). This is not the case, however (see also Kennedy, 2001). The believer in precognition may object that psychic abilities, unlike all other abilities, are not influenced by natural selection. But the onus is then squarely on the believer in psi to explain why this should be so.

Second, there is no real-life evidence that people can feel the future (e.g., nobody has ever collected the $1 million available for anybody who can demonstrate paranormal performance under controlled conditions).[4] To appreciate how unlikely the existence of psi really is, consider the facts that (a) casinos make profits and (b) casinos feature the game of French roulette. French roulette has 37 numbers, 18 colored black, 18 colored red, and the special number 0. The situation we consider here is where gamblers bet on the color indicated by the roulette ball. Betting on the wrong color results in a loss of your stake, and betting on the right color will double your stake. Because of the special number 0, the house holds a small advantage over the gambler; the probability of the house winning is 19/37.

Consider now the possibility that the gambler could use psi to bet on the color that will shortly come up (i.e., the color that will bring great wealth in the immediate future). In this context, even small effects of psi result in substantial payoffs. For instance, suppose a player with psi can anticipate the correct color in 53.1% of cases—the mean percentage correct across participants for the erotic pictures in Bem's (2011) Experiment 1. Assume that this psi player starts with only 100 euros and bets 10 euros every time. The gambling stops whenever the psi player is out of money (in which case the casino wins) or the psi player has accumulated €1 million. After accounting for the house advantage, what is the probability that the psi player will win €1 million? This probability, easily calculated from random walk theory (e.g., Feller, 1970, 1971), equals 48.6%. This means that, in this case, the expected profit for a psychic's night out at the casino equals $485,900. If Bem's psychic plays the game all year round, never raises the stakes, and always quits at a profit of a million dollars, the expected return is $177,353,500.[5]

Clearly, Bem's psychic could bankrupt all casinos on the planet before anybody realized what was going on. This analysis leaves us with two possibilities. The first possibility is that, for whatever reason, the psi effects are not operative in casinos, but they are operative in psychological experiments on erotic pictures. The second possibility is that the psi effects are either nonexistent or else so small that they cannot overcome the house advantage. Note that in the latter case, all of Bem's (2011) experiments overestimated the effect.

To return to Laplace's principle, we feel the above reasons motivate us to assign our prior belief in precognition a number very close to zero. For illustrative purposes, let us set $P(H_1) = 10^{-20}$, that is, .00000000000000000001. This means that $P(H_0) = 1 - P(H_1) = .99999999999999999999$. Our aim here is not to quantify precisely our personal prior belief in psi. Instead, our aim is to explain Laplace's principle by using a concrete example and specific numbers. It is also important to note that the Bayesian $t$ test outlined in the next section does not depend in any way on the prior probabilities $P(H_0)$ and $P(H_1)$.

Now assume we find a flawless, well-designed, 100% confirmatory experiment for which the observed data are unlikely under $H_0$ but likely under $H_1$, say by a factor of 19 (as indicated below, this is considered strong evidence). In order to update our prior belief, we apply Bayes' rule:

$$\begin{aligned} p(H_1|D) &= \frac{p(D|H_1)p(H_1)}{p(D|H_0)p(H_0) + p(D|H_1)p(H_1)} \\ &= \frac{.95 \times 10^{-20}}{.05(1 - 10^{-20}) + .95 \times 10^{-20}} \\ &= .00000000000000000019. \end{aligned}$$

[3] This is evident from the fact that psi research is almost never published in the mainstream literature.

[4] See http://www.skepdic.com/randi.html for details.

[5] The break-even point for the house lies at a success probability of .514. However, even if the success rate is smaller, say, .510, one can boost one's success probability by utilizing a team of psychics and using their majority vote. This is so because Condorcet's jury theorem ensures that, whenever the success probability for an individual voter lies above .5, the probability of a correct majority vote approaches 1 as the number of voters grows large. If the individual success probability is .510, for instance, using the majority vote of a team of 1,000 psychics gives a probability of .73 for the majority vote being correct.

True, our posterior belief in precognition is now higher than our prior belief. Nevertheless, we are still relatively certain that precognition does not exist. In order to overcome our skeptical prior opinion, the evidence must be much stronger. In other words, extraordinary claims require extraordinary evidence. This is neither irrational nor unfair; if the proponents of precognition succeed in establishing its presence, their reward is eternal fame (and, if Bem were to take his participants to the casino, infinite wealth).

Thus, in order to convince scientific critics of an extravagant or controversial claim, one is required to pull out all the stops. Even if Bem's (2011) experiments had been confirmatory (which they were not, see above), and even if they had conveyed strong statistical evidence for precognition (which they did not, see below), eight experiments are not enough to convince a skeptic that the known laws of nature have been bent. Or, more precisely, that these laws were bent only for erotic pictures and only for participants who are extraverts.

## Problem 3: *p* Values Overstate the Evidence Against the Null

Consider a data set for which $p = .001$, indicating a low probability of encountering a test statistic that is at least as extreme as the one that was actually observed, given that the null hypothesis $H_0$ is true. Should we proceed to reject $H_0$? Well, this depends at least in part on how likely the data are under $H_1$. Suppose, for instance, that $H_1$ represents a very small effect. Then it may be that the observed value of the test statistic is almost as unlikely under $H_0$ as under $H_1$. What is going on here?

The underlying problem is that evidence is a relative concept, and it is of limited interest to consider the probability of the data under just a single hypothesis. For instance, if you win the state lottery you might be accused of cheating; after all, the probability of winning the state lottery is rather small. This may be true, but this low probability in itself does not constitute evidence. The evidence is assessed only when this low probability is pitted against the much lower probability that you could somehow have obtained the winning number by acquiring advance knowledge on how to buy the winning ticket.

Therefore, in order to evaluate the strength of evidence that the data provide for or against precognition, we need to pit the null hypothesis against a specific alternative hypothesis and not consider the null hypothesis in isolation. Several methods are available to achieve this goal. Classical statisticians can achieve this goal with the Neyman–Pearson procedure; statisticians who focus on likelihood can achieve this goal using likelihood ratios (Royall, 1997); and Bayesian statisticians can achieve this goal using a hypothesis test that computes a weighted likelihood ratio (e.g., Rouder et al., 2009; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009). As an illustration, we focus here on the Bayesian hypothesis test.

In a Bayesian hypothesis test, the goal is to quantify the change in prior to posterior odds that is brought about by the data. For a choice between $H_0$ and $H_1$, we have

$$\frac{p(H_0|D)}{p(H_1|D)} = \frac{p(H_0)}{p(H_1)} \times \frac{p(D|H_0)}{p(D|H_1)}, \quad (1)$$

which is often verbalized as

$$\text{Posterior model odds} = \text{Prior model odds} \times \text{Bayes factor}. \quad (2)$$

Thus, the change from prior odds $p(H_0)/p(H_1)$ to posterior odds $p(H_0|D)/p(H_1|D)$ brought about by the data is given by the ratio of $p(D|H_0)/p(D|H_1)$, a quantity known as the *Bayes factor* (Jeffreys, 1961). The Bayes factor (or its logarithm) is often interpreted as the weight of evidence provided by the data (Good, 1985; for details, see Berger & Pericchi, 1996; Bernardo & Smith, 1994, Chapter 6, Gill, 2002, Chapter 7, Kass & Raftery, 1995; O'Hagan, 1995).

When the Bayes factor for $H_0$ over $H_1$ equals 2 (i.e., $BF_{01} = 2$), this indicates that the data are twice as likely to have occurred under $H_0$ than under $H_1$. Even though the Bayes factor has an unambiguous and continuous scale, it is sometimes useful to summarize the Bayes factor in terms of discrete categories of evidential strength. Jefferys (1961, Appendix B) proposed the classification scheme shown in Table 1.

Several researchers have recommended Bayesian hypothesis tests (e.g., Berger & Delampady, 1987; Berger & Sellke, 1987; Edwards, Lindman, & Savage, 1963; see also Wagenmakers & Grünwald, 2006), particularly in the context of psi (e.g., Bayarri & Berger, 1991; Jaynes, 2003, Chapter 5; Jefferys, 1990).

To illustrate the extent to which Bem's (2011) conclusions depend on the statistical test that was used, we have reanalyzed the Bem experiments with a default Bayesian *t* test (Gönen et al., 2005; Rouder et al., 2009). This test computes the Bayes factor for $H_0$ versus $H_1$, and it is important to note that the prior model odds plays no role whatsoever in its calculation (see also Equations 1 and 2). One of the advantages of this Bayesian test is that it also allows researchers to quantify the evidence in favor of the null hypothesis, something that is impossible with traditional *p* values. Another advantage of the Bayesian test is that it is *consistent*: As the number of participants grows large, the probability of discovering the true hypothesis approaches 1.

## The Bayesian *t* Test

Ignoring for the moment our concerns about the exploratory nature of the Bem (2011) studies and the prior odds in favor of the

Table 1
*Classification Scheme for the Bayes Factor, as Proposed by Jeffreys (1961)*

| Bayes factor, $BF_{01}$ | Interpretation |
|---|---|
| >100 | Extreme evidence for $H_0$ |
| 30–100 | Very strong evidence for $H_0$ |
| 10–30 | Strong evidence for $H_0$ |
| 3–10 | Substantial evidence for $H_0$ |
| 1–3 | Anecdotal evidence for $H_0$ |
| 1 | No evidence |
| 1/3–1 | Anecdotal evidence for $H_1$ |
| 1/10–1/3 | Substantial evidence for $H_1$ |
| 1/30–1/10 | Strong evidence for $H_1$ |
| 1/100–1/30 | Very strong evidence for $H_1$ |
| <1/100 | Extreme evidence for $H_1$ |

*Note.* We replaced the labels "worth no more than a bare mention" with "anecdotal" and "decisive" with "extreme."

null hypothesis, we can wonder how convincing the statistical results from the Bem studies really are. After all, each of the Bem studies featured at least 100 participants, but nonetheless in several experiments Bem had to report one-sided (not two-sided) $p$ values in order to claim significance at the .05 level. One might intuit that such data do not constitute compelling evidence for precognition.

In order to assess the strength of evidence for $H_0$ (i.e., no precognition) versus $H_1$ (i.e., precognition), we computed a default Bayesian $t$ test for the critical tests reported in Bem (2011). This default test is based on general considerations that represent a lack of knowledge about the effect size under study (Gönen et al., 2005; Rouder et al., 2009; for a generalization to regression, see Liang, Paulo, Molina, Clyde, & Berger, 2008). More specific assumptions about the effect size of psi would result in a different test. We decided to first apply the default test because we did not feel qualified to make these more specific assumptions, especially not in an area as contentious as psi.

With the Bayesian $t$-test web applet provided by J. N. Rouder,[6] it is straightforward to compute the Bayes factor for the Bem experiments: All that is needed is the $t$ value and the degrees of freedom (Rouder et al., 2009). Table 2 shows the results. Out of the 10 critical tests, only one yields "substantial" evidence for $H_1$, whereas three yield "substantial" evidence in favor of $H_0$. The results of the remaining six tests provide evidence that is only "anecdotal" or "worth no more than a bare mention" (Jeffreys, 1961).

In sum, a default Bayesian test confirms the intuition that, for large sample sizes, one-sided $p$ values higher than .01 are not compelling (see also Wetzels et al., in press).[7] Overall, the Bayesian $t$ test indicates that the data of Bem (2011) do not support the hypothesis of precognition. This is despite the fact that multiple hypotheses were tested, something that warrants a correction (for a Bayesian correction, see Scott & Berger, 2010; Stephens & Balding, 2009).

Note that, even though our analysis is Bayesian, we did not select priors to obtain a desired result: The Bayes factors that were calculated are independent of the prior model odds and depend only on the prior distribution for effect size; for this distribution, we used the default option. We also examined other options, however, and found that our conclusions are robust: For a wide range of different, nondefault prior distributions on effect size, the evidence for precognition is either nonexistent or negligible.[8]

Table 2
*Results of 10 Crucial Tests for the Experiments Reported in Bem (2011), Reanalyzed With the Default Bayesian* t *Test*

| Experiment | *df* | $\|t\|$ | *p* | $BF_{01}$ | Evidence category (in favor of *H.*) |
|---|---|---|---|---|---|
| 1 | 99 | 2.51 | 0.01 | 0.61 | Anecdotal ($H_1$) |
| 2 | 149 | 2.39 | 0.009 | 0.95 | Anecdotal ($H_1$) |
| 3 | 96 | 2.55 | 0.006 | 0.55 | Anecdotal ($H_1$) |
| 4 | 98 | 2.03 | 0.023 | 1.71 | Anecdotal ($H_0$) |
| 5 | 99 | 2.23 | 0.014 | 1.14 | Anecdotal ($H_0$) |
| 6 | 149 | 1.80 | 0.037 | 3.14 | Substantial ($H_0$) |
| 6 | 149 | 1.74 | 0.041 | 3.49 | Substantial ($H_0$) |
| 7 | 199 | 1.31 | 0.096 | 7.61 | Substantial ($H_0$) |
| 8 | 99 | 1.92 | 0.029 | 2.11 | Anecdotal ($H_0$) |
| 9 | 49 | 2.96 | 0.002 | 0.17 | Substantial ($H_1$) |

*Note.* *df* = degrees of freedom; $BF_{01}$ = Bayes factor; $H_0$ = precognition does not exist; $H_1$ = precognition does exist.

At this point, one may wonder whether it is feasible to use the Bayesian $t$ test and eventually obtain enough evidence against the null hypothesis to overcome the prior skepticism outlined in the previous section. Indeed, this is feasible: Based on the mean and sample standard deviations reported in Bem's (2011) Experiment 1, it is straightforward to calculate that around 2,000 participants are sufficient to generate an extremely high Bayes factor $BF_{01}$ of about $10^{-24}$; when this extreme evidence is combined with the skeptical prior, the end result is firm belief that psi is indeed possible. On the one hand, 2,000 participants seems excessive; on the other hand, this is but a small subset of participants who have been tested in the field of parapsychology during the last decade. Of course, this presupposes that the experiment under consideration was 100% confirmatory and that it was conducted with the utmost care.

## Guidelines for Confirmatory Research

As discussed earlier, exploratory research is useful but insufficiently compelling to change the mind of a skeptic. In order to provide hard evidence for or against an empirical proposition, one has to resort to strictly confirmatory studies. The degree to which the scientific community will accept semiconfirmatory studies as evidence depends partly on the plausibility of the claim under scrutiny. Again, extraordinary claims require extraordinary evidence. The basic characteristic of confirmatory studies is that all choices that could influence the result have been made before the data are observed. We suggest that confirmatory research in psychology observes the following guidelines:

1. Fishing expeditions should be prevented by selecting participants and items before the confirmatory study takes place. Of course, previous tests, experiments, and questionnaires may be used to identify those participants and items that show the largest effects. This method increases power in case the phenomenon of interest really does exist; however, no further selection or subset testing should take place once the confirmatory experiment has started.

2. Data should be transformed only if this has been decided beforehand. In confirmatory studies, one does not "torture the data until they confess." It also means that, upon failure, confirmatory experiments are not demoted to exploratory pilot experiments, and that, upon success, exploratory pilot experiments are not promoted to confirmatory experiments.

3. In simple examples, such as when the dependent variable is success rate or mean response time, an appropriate analysis should be decided upon before the data have been collected.

4. It is prudent to report more than a single statistical analysis. If the conclusions from $p$ values conflict with those of, say, Bayes factors, this should be clearly stated. Compelling results yield similar conclusions, irrespective of the statistical paradigm that is used to analyze the data.

[6] See http://pcl.missouri.edu/bayesfactor

[7] A preprint is available at http://www.ruudwetzels.com/

[8] This robustness analysis is reported in an online appendix available on the first author's website (http://www.ejwagenmakers.com/papers.html).

In our opinion, the above guidelines are sufficient for most research topics. However, the researcher who wants to convince a skeptical community of academics that psi exists may want to go much further. In the context of psi, Price (1955) argued that "what is needed is something that can be demonstrated to the most hostile, pig-headed, and skeptical of critics" (p. 365). This is also consistent with Hume's maxim that "no testimony is sufficient to establish a miracle, unless the testimony be of such a kind, that its falsehood would be more miraculous, than the fact, which it endeavours to establish" (Hume, 1748/1910, Chapter 10). What this means is that in order to overcome the skeptical bias against psi, the psi researcher might want to consider more drastic measures to ensure that the experiment was completely confirmatory:

5. The psi researcher may make stimulus materials, computer code, and raw data files publicly available online. The psi researcher may also make the decisions made with respect to Guidelines 1–4 publicly available online and do so before the confirmatory experiment is carried out.

6. The psi researcher may engage in an adversarial collaboration, that is, a collaboration with a true skeptic, and preferably more than one (Price, 1955; Wiseman & Schlitz, 1997). This echoes the advice of Diaconis (1991), who stated that the studies on psi reviewed by (Utts, 1991) were "crucially flawed. . . . Since the field has so far failed to produce a replicable phenomena, it seems to me that any trial that asks us to take its findings seriously should include full participation by qualified skeptics" (p. 386).

The psi researcher who also follows the last two guidelines makes an effort that is slightly higher than usual; we believe this is a small price to pay for a large increase in credibility. It should after all be straightforward to document the intended analyses, and in most universities a qualified skeptic is sitting in the office next door.

## Concluding Comment

In eight out of nine studies, Bem (2011) reported evidence in favor of precognition. As we have argued above, this evidence may well be illusory; in several experiments, it is evident that exploration should have resulted in a correction of the statistical results. Also, we have provided an alternative, Bayesian reanalysis of Bem's experiments; this alternative analysis demonstrated that the statistical evidence was, if anything, slightly in favor of the null hypothesis. One can argue about the relative merits of classical *t* tests versus Bayesian *t* tests, but this is not our goal; instead, we want to point out that the two tests yield very different conclusions, something that casts doubt on the conclusiveness of the statistical findings.

In this article, we have assessed the evidential impact of Bem's (2011) 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) 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 change of $10^{1000}$ to 1—but just one good experiment" (p. 367).

Although the Bem (2011) experiments themselves do not provide evidence for precognition, they do suggest that our academic standards of evidence may currently be set at a level that is too low (see also Wetzels et al., in press). 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-à-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 (see also Ioannidis, 2005). 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.

## References


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

Bem, D. J. (2000). Writing an empirical article. In R. J. Sternberg (Ed.), *Guide to publishing in psychology journals* (pp. 3–16). Cambridge, England: Cambridge University Press.

Bem, D. J. (2003). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger III (Eds.), *The compleat academic: A career guide* (pp. 171–201). Washington, DC: American Psychological Association.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100,* 1–19.

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

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

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p* values and evidence. *Journal of the American Statistical Association, 82,* 112–139.

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

Diaconis, P. (1978, July 14). Statistical problems in ESP research. *Science, 201,* 131–136.

Diaconis, P. (1991). Comment. *Statistical Science, 6,* 386.

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

Feller, W. (1970). *An introduction to probability theory and its applications* (Vol. 1). New York, NY: Wiley.

Feller, W. (1971). *An introduction to probability theory and its applications* (Vol. 2). New York, NY: Wiley.

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

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

Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo,

M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 2* (pp. 249–269). New York, NY: Elsevier.

Hume, D. (1910). *An enquiry concerning human understanding.* Retrieved from http://www.scribd.com/doc/413830.David-Hume-An-Enquiry-Concerning Human-Understanding (Original work published 1748).

Hyman, R. (2007). Evaluating parapsychological claims. In R. J. Sternberg, H. L. Roediger III, & D. F. Halpern (Eds.), *Critical thinking in psychology* (pp. 216–231). Cambridge, England: Cambridge University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* 696–701.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge, England: Cambridge University Press.

Jefferys, W. H. (1990). Bayesian analysis of random event generator data. *Journal of Scientific Exploration, 4,* 153–169.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Kennedy, J. E. (2001). Why is psi so elusive? A review and proposed model. *Journal of Parapsychology, 65,* 219–246.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217.

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.

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

Price, G. R. (1955). Science and the supernatural. *Science, 122,* 359–367.

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.

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm.* London, England: Chapman & Hall.

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics, 38,* 2587–2619.

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics, 10,* 681–690.

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. *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.

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.

Wiseman, R., & Schlitz, M. (1997). Experimenter effects and the remote detection of staring. *Journal of Parapsychology, 61,* 197–207.

Received October 18, 2010
Revision received December 23, 2010
Accepted January 7, 2011 ■