Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi

Eric–Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, & Han van der Maas

University of Amsterdam

Abstract

Does psi exist? In a recent article, Dr. Bem conducted nine studies with over a thousand 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 using 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.

In a recent article for *Journal of Personality and Social Psychology*, Bem (in press) presented nine experiments that test for the presence of psi.[1] Specifically, 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 that cannot presently be explained in terms of a known biological or physical mechanism.

Despite the lack of a plausible mechanistic account of precognition, Bem was able to reject the null hypothesis of no precognition in eight out of nine experiments. For instance,

---

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

---

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 takes 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". 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 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—WWBM], 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 experiments, discussed below in detail, are the following: (1) confusion between exploratory and confirmatory studies, brought about by what we have termed the Bem Exploration Method (BEM); (2) 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); (3) 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, 2009) to quantify the evidence that Bem (in press) presents 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. 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 danger to the field, and impedes progress on phenomena that are replicable and important.

Problem 1: The Bem Exploration Method

In his popular book chapter "Writing the empirical journal article", Bem provides the following advice to graduate students:

"There are two possible articles you can write: (1) the article you planned
to write when you designed your study or (2) the article that makes the most

sense now that you have seen the results. They are rarely the same, and the correct answer is (2)." (Bem, 2003, pp. 171-172)

We coin this strategy the Bem Exploration Method (BEM). Clearly, the method implicitly holds that all research should be explorative research. We agree with Bem that exploration is important, but would insist that whether a study is exploratory or not should always be clearly indicated in a paper. In particular, testing multiple hypotheses to subsequently report the best results as if they had come about through a confirmatory study is at odds with the basic ideas underlying scientific methodology.

Bem continues:

"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. If this is how our enterprise actually proceeded, we could write most of the article before we collected the data. We could write the introduction and method sections completely, prepare the results section in skeleton form, leaving spaces to be filled in by the specific numerical results, and have two possible discussion sections ready to go, one for positive results, the other for negative results. But this is not how our enterprise actually proceeds. Psychology is more exciting than that (...)" (Bem, 2003, p. 172).

This may be true, but if one wants to convince a skeptical audience, as in the case of psi, a confirmatory study is much more compelling than an exploratory study. 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 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 non-erotic. 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. The Bem Exploration Method holds that a new and different story would then have been constructed around these other results. This means that Bem's Experiment 1 was to some extent a fishing expedition, an expedition that should have resulted in a correction of the reported p-value.

Another example of exploration comes from Bem's 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 Experiment 5 the analysis shows that "Women achieved a significant hit rate on the negative pictures, 53.6%, $t(62) = 2.25$, $p = .014$, $d = .28$; but men did not, 52.4%, $t(36) = 0.89$, $p = .19$, $d = .15$." But why test for gender in the first place? There appears to be no good reason. Indeed, Bem himself states that "the psi literature does not reveal any systematic sex differences in psi ability".

Bem's 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." Again, conducting multiple tests requires a correction.

These explorative elements are clear from Bem's discussion of the empirical data. The problem with Bem's BEM 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 upwards.

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 they were 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 they are 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 goldfish can talk, that the earth will perish in 2012, and 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, and should be, very low. First, there exists no mechanistic theory of precognition (see Price, 1955 for a discussion). 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, in press), and one might therefore assume that natural selection would have lead to a world filled with powerful psychics (i.e., people or animals with precognition, clairvoyance, psychokineses, etc.). 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 $1,000,000 available for anybody who can demonstrate paranormal

performance under controlled conditions[2], etc.). To appreciate how unlikely the existence of psi really is, consider the facts that (a) casinos make profit, and (b) casinos feature the game of French roulette. French roulette features 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, that is, 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 Experiment 1. Assume that this psi-player starts with only 100 euros, and bets 10 euro 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 one million euros. After accounting for the house advantage, what is the probability that the psi-player will win one million euros? 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.

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 experiments overestimate the effect.

Returning to Laplace's Principle, we should obviously assign our prior belief in precognition a number very close to zero, perhaps slightly larger than the probability of, say, goldfish being able to talk. For illustrative purposes, let us set $P(H_1) = 10^{-20}$, that is, .00000000000000000001. This means that $P(H_0) = 1 - P(H_1) = .99999999999999999999$.

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:

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

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

---

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

skeptical prior opinion, the evidence needs to 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 when Bem's experiments had been confirmatory (which they were not, see above), and even if they would have 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 not insightful 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

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

which is often verbalized as

$$\text{Posterior model odds} = \text{Prior model odds} \times \text{Bayes factor.} \tag{2}$$

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, and 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 then 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. Jeffreys (1961, Appendix B) proposed the classification scheme shown in Table 1.

Table 1: Classification scheme for the Bayes factor, as proposed by Jeffreys (1961). We replaced the labels "worth no more than a bare mention" with "anecdotal", and "decisive" with "extreme".

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

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, Chap. 5; Jeffreys, 1990).

To illustrate the extent to which Bem's 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 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 studies, and the prior odds in favor of the 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

Table 2: The results of 10 crucial tests for the experiments reported in Bem (in press), reanalyzed using the default Bayesian t-test.

Exp	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)

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 (in press). 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 apply the default test because we do not feel qualified to make these more specific assumptions, especially not in an area as contentious as psi.

Using the Bayesian t-test web applet provided by Dr. Rouder[3] 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. Overall, the Bayesian t-test indicates that the data of Bem 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. Thus, the foregoing shows that there exists a reasonable default test according to which the Bem experiments yield no evidence for precognition.

---

[3]See `http://pcl.missouri.edu/bayesfactor`.

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 Experiment 1, it is straightforward to calculate that around 2000 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, 2000 participants seems excessive; on the other hand, this is but a small subset of participants that 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 has been conducted with the utmost care.

Six Guidelines for Research on Psi

How should research on psi proceed in order to be immune to most criticism? As argued by Price (1955, p. 365), "(...) what is needed is something that can be demonstrated to the most hostile, pig-headed, and skeptical of critics." In order to achieve this aim, we propose that at least the following requirements need to be fulfilled:

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 who show the largest effects—this method increases power in case psi really does exist; however, no further selection or subset testing should take place once the confirmatory experiment has started.

2. 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. Because evidence is a relative concept, such an analysis should quantify evidence for H_0 versus H_1. When the researcher chooses to compute a Bayes factor, this may be done using default priors, or different priors based on expertise and experience (Goldstein, 2006); however, such informative priors need to be formulated *before* the confirmatory experiment has started.

3. In order to ensure that steps 1 and 2 are followed to the letter, we recommend that the psi researcher engages in an adversarial collaboration, that is, collaboration with a true skeptic, and preferably more than one (Diaconis, 1991; Price, 1955; Wiseman & Schlitz, 1997).

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, then the results are probably not compelling. Compelling results yield similar conclusions, irrespective of the statistical paradigm that is used to analyze the data.

5. Because exploratory analyses need to be avoided as much as possible, participants are excluded only for reasons that have been articulated explicitly before the confirmatory experiment takes place; in the same spirit, data should only be transformed unless this

has been decided beforehand. 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.

6. The stimulus materials, computer code, and raw data files for all participants should be made publicly available online. We also recommend that the decisions made in step 1 and 2 are made publicly available online *before* the confirmatory experiment is conducted. This procedure will hopefully counteract, at least to some extent, the file drawer problem.

In the context of research on precognition and psi, the above requirements are only sensible, and psi researchers who wish to convince the academic world that the phenomenon exists are well advised to heed them (see also Price, 1955). Note that none of these requirements were fulfilled by the Bem experiments.

Concluding Comment

In eight out of nine studies, Bem reported evidence in favor of precognition. As we have argued above, this evidence may well be illusory; in several experiments it is evident that Bem's Exploration Method 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.

Although the Bem 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. 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.

References

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

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. (in press). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology.*

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). 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. I.* New York: John Wiley & Sons.

Feller, W. (1971). *An introduction to probability theory and its applications: Vol. ii.* New York: John Wiley & Sons.

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

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis, 1*, 403-420.

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two–sample *t* test. *The 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: Elsevier.

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: Cambridge University Press.

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

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

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

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. *The Journal of Parapsychology, 65*, 219–246.

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 B*, *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: Chapman & Hall.

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical–Bayes multiplicity adjustment in the variable–selection problem. *The 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.

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.

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.