Check for updates

RESEARCH ARTICLE

Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events [version 1; referees: 2 approved]

Daryl Bem[1], Patrizio Tressoldi[2], Thomas Rabeyron[3,4], Michael Duggan[5]

[1]Cornell University, New York, NY, 10011, USA
[2]Università di Padova, Padova, 35122, Italy
[3]Université de Nantes, Nantes, 44300, France
[4]University of Edinburgh, Edinburgh, Scotland, EH8 9YL, UK
[5]Nottingham Trent University, Nottingham, England, NG1 4BU, UK

v1 **First published:** 30 Oct 2015, **4**:1188 (doi: 10.12688/f1000research.7177.1)
Latest published: 29 Jan 2016, **4**:1188 (doi: 10.12688/f1000research.7177.2)

Abstract

In 2011, one of the authors (DJB) published a report of nine experiments in the *Journal of Personality and Social Psychology* purporting to demonstrate that an individual's cognitive and affective responses can be influenced by randomly selected stimulus events that do not occur until after his or her responses have already been made and recorded, a generalized variant of the phenomenon traditionally denoted by the term *precognition*. To encourage replications, all materials needed to conduct them were made available on request. We here report a meta-analysis of 90 experiments from 33 laboratories in 14 countries which yielded an overall effect greater than 6 sigma, $z = 6.40$, $p = 1.2 \times 10^{-10}$ with an effect size (Hedges' g) of 0.09. A Bayesian analysis yielded a Bayes Factor of 1.4×10^{9}, greatly exceeding the criterion value of 100 for "decisive evidence" in support of the experimental hypothesis. When DJB's original experiments are excluded, the combined effect size for replications by independent investigators is 0.06, $z = 4.16$, $p = 1.1 \times 10^{-5}$, and the BF value is 3,853, again exceeding the criterion for "decisive evidence." The number of potentially unretrieved experiments required to reduce the overall effect size of the complete database to a trivial value of 0.01 is 544, and seven of eight additional statistical tests support the conclusion that the database is not significantly compromised by either selection bias or by "*p*-hacking"—the selective suppression of findings or analyses that failed to yield statistical significance. *P*-curve analysis, a recently introduced statistical technique, estimates the true effect size of our database to be 0.20, virtually identical to the effect size of DJB's original experiments (0.22) and the closely related "presentiment" experiments (0.21). We discuss the controversial status of precognition and other anomalous effects collectively known as *psi*.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	 2
REVISED **version 2** published 29 Jan 2016		
version 1 published 30 Oct 2015	✓ report	✓ report

1 **Ina Vitalevna Vasileva**, Tyumen State University Russian Federation

2 **Paul Grigoriev**, Crimea State Medical University Ukraine

Discuss this article

Comments (3)

Corresponding author: Daryl Bem (d.bem@cornell.edu)

How to cite this article: Bem D, Tressoldi P, Rabeyron T and Duggan M. **Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events [version 1; referees: 2 approved]** *F1000Research* 2015, **4**:1188 (doi: 10.12688/f1000research.7177.1)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: No competing interests were disclosed.

First published: 30 Oct 2015, **4**:1188 (doi: 10.12688/f1000research.7177.1)

In 2011, the *Journal of Personality and Social Psychology* published an article by one of us (DJB) entitled "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect" (Bem, 2011). The article reported nine experiments that purported to demonstrate that an individual's cognitive and affective responses can be influenced by randomly selected stimulus events that do not occur until after his or her responses have already been made and recorded, a generalized variant of the phenomenon traditionally denoted by the term *precognition*. The controversial nature of these findings prompted the journal's editors to publish an accompanying editorial justifying their decision to publish the report and expressing their hope and expectation that attempts at replication by other investigators would follow (Judd & Gawronski, 2011).

To encourage replications from the beginning of his research program in 2000, Bem offered free, comprehensive packages that included detailed instruction manuals for conducting the experiments, computer software for running the experimental sessions, and database programs for collecting and analyzing the data. As of September 2013, two years after the publication of his article, we were able to retrieve 69 attempted replications of his experiments and 11 other experiments that tested for the anomalous anticipation of future events in alternative ways. When Bem's experiments are included, the complete database comprises 90 experiments from 33 different laboratories located in 14 different countries.

Precognition is one of several phenomena in which individuals appear to have access to "nonlocal" information, that is, to information that would not normally be available to them through any currently known physical or biological process. These phenomena, collectively referred to as *psi*, include *telepathy,* access to another person's thoughts without the mediation of any known channel of sensory communication; *clairvoyance* (including a variant called *remote viewing*), the apparent perception of objects or events that do not provide a stimulus to the known senses; and *precognition*, the anticipation of future events that could not otherwise be anticipated through any known inferential process.

Laboratory-based tests of precognition have been published for nearly a century. Most of the earlier experiments used forced-choice designs in which participants were explicitly challenged to guess on each trial which one of several potential targets would be randomly selected and displayed in the near future. Typical targets included ESP card symbols, an array of colored light bulbs, the faces of a die, or visual elements in a computer display. When a participant correctly predicted the actual target-to-be, the trial was scored as a hit, and performance was typically expressed as the percentage of hits over a given number of trials.

A meta-analysis of all forced-choice precognition experiments appearing in English language journals between 1935 and 1977 was published by Honorton & Ferrari (1989). Their analysis included 309 experiments conducted by 62 different investigators involving more than 50,000 participants. Honorton and Ferrari reported a small but significant hit rate, Rosenthal effect size $z/\sqrt{n} = .02$, Stouffer $Z = 6.02$, $p = 1.1 \times 10^{-9}$. They concluded that this overall result was unlikely to be artifactually inflated by the selective reporting of positive results (the so-called file-drawer effect), calculating that there would have to be 46 unreported studies averaging null results for every reported study in the meta-analysis to reduce the overall significance of the database to chance.

Just as research in cognitive and social psychology has increasingly pursued the study of affective and cognitive processes that are not accessible to conscious awareness or control (e.g., Ferguson & Zayas, 2009), research in psi has followed the same path, moving from explicit forced-choice guessing tasks to experiments using subliminal stimuli and implicit or physiological responses. This trend is exemplified by several "presentiment" experiments, pioneered by Radin (1997) and Bierman (Bierman & Radin, 1997) in which physiological indices of participants' emotional arousal are continuously monitored as they view a series of pictures on a computer screen. Most of the pictures are emotionally neutral, but on randomly selected trials, a highly arousing erotic or negative image is displayed. As expected, participants show strong physiological arousal when these images appear, but the important "presentiment" finding is that the arousal is observed to occur a few seconds before the picture actually appears on the screen—even before the computer has randomly selected the picture to be displayed.

The presentiment effect has now been demonstrated using a variety of physiological indices, including electrodermal activity, heart rate, blood volume, pupil dilation, electroencephalographic activity, and fMRI measures of brain activity. A meta-analysis of 26 reports of presentiment experiments published between 1978 and 2010 yielded an average effect size of 0.21, 95% CI = [0.13, 0.29], combined $z = 5.30$, $p = 5.7 \times 10^{-8}$. The number of unretrieved experiments averaging a null effect that would be required to reduce the effect size to a trivial level was conservatively calculated to be 87 (Mossbridge *et al.*, 2012; see also, Mossbridge *et al.*, 2014). A critique of this meta-analysis has been published by Schwarzkopf (2014) and the authors have responded to that critique (Mossbridge *et al.*, 2015).

Bem's experiments can be viewed as direct descendants of the presentiment experiments. Like them, each of his experiments modified a well-established psychological effect by reversing the usual time-sequence of events so that the participant's responses were obtained before the putatively causal stimulus events occurred. The hypothesis in each case was that the time-reversed version of the experiment would produce the same result as the standard non-time-reversed experiment. Four well-established psychological effects were modified in this way. (See Bem (2011) for more complete descriptions of the experimental protocols.)

Precognitive approach and avoidance

Two experiments tested time-reversed versions of one of psychology's oldest and best known phenomena, the Law of Effect (Thorndike, 1898): An organism is more likely to repeat responses that have been positively reinforced in the past than responses that have not been reinforced. Bem's time-reversed version of this effect tested whether participants were more likely to make responses that would be reinforced in the near future. On each trial of the first experiment ("Precognitive Detection of Erotic Stimuli"), the participant selected one of two curtains displayed side-by-side on a computer screen. After the participant had made a choice, the computer randomly designated one of the curtains to be the reinforced alternative. If the participant had selected that curtain, it opened to reveal an erotic photograph and the trial was scored as a hit; if the participant had selected the other curtain, a blank gray wall appeared and the trial was scored as a miss. In a second experiment ("Precognitive Avoidance of Negative Stimuli") a trial was scored as a hit if the participant selected the alternative that avoided the display of a gruesome or unpleasant photograph.

Retroactive priming

In recent years, priming experiments have become a staple of cognitive social psychology (Klauer & Musch, 2003). In a typical affective priming experiment, participants are asked to judge as quickly as they can whether a photograph is pleasant or unpleasant and their response time is measured. Just before the picture appears, a positive or negative word (e.g., *beautiful*, *ugly*) is flashed briefly on the screen; this word is called the prime. Individuals typically respond more quickly when the valences of the prime and the photograph are congruent (both are positive or both are negative) than when they are incongruent. In the time-reversed version of the procedure, the randomly-selected prime appeared after rather than before participants judge the affective valence of the photograph.

Retroactive habituation

When individuals are initially exposed to an emotionally arousing stimulus, they typically have a strong physiological response to it. Upon repeated exposures the arousal diminishes. This *habituation* process is one possible mechanism behind the so-called "mere exposure" effect in which repeated exposures to a stimulus produce increased liking for it (Bornstein, 1989; Zajonc, 1968). It has been suggested that if a stimulus is initially frightening or unpleasant, repeated exposures will render it less negatively arousing and, hence, it will be better liked after the exposures—the usual mere exposure result—but if the stimulus is initially very positive, the repeated exposures will render it boring or less positively arousing and, hence, it will be *less* well liked after the exposures (Dijksterhuis & Smith, 2002).

In two time-reversed habituation experiments, pairs of negative photographs matched for equal likeability or pairs of erotic photographs similarly matched were displayed side by side on the screen and the participant was instructed on each trial to indicate which one he or she liked better. After the preference was recorded, the computer randomly selected one of the two photographs to be the habituation target and flashed it subliminally on the screen several times. The hypothesis was that participants would prefer the habituation target on trials with negative photographs but would prefer the nontarget on trials with erotic photographs.

The three time-reversed effects described above can be viewed as conceptual replications of the presentiment experiments in that all these experiments assessed affective responses to emotionally arousing stimuli before those stimuli were randomly selected and displayed. Whereas presentiment experiments assess physiological responses, Bem's experiments assessed behavioral responses. Even the photographs used in the two kinds of experiments were drawn primarily from the same source, the International Affective Picture System (IAPS; Lang & Greenwald, 1993), a set of more than 800 digitized photographs that have been rated for valence and arousal.

Retroactive facilitation of recall

A commonplace phenomenon of memory is that practicing or rehearsing a set of verbal items facilitates their subsequent recall. Two of Bem's time-reversed experiments tested whether rehearsing a set of words makes them easier to recall even if the rehearsal takes place after the recall test is administered. Participants were shown 48 common nouns one at a time on the computer screen. They were then given a (surprise) recall test in which they were asked to type out all the words they could recall, in any order. After the participant completed the recall test, the computer randomly selected half the words to serve as practice words and had participants rehearse them in a series of practice exercises. The hypothesis was that this practice would "reach back in time" to facilitate the recall of these words and, thus, participants would recall more of the to-be-practiced words than the control non-practiced words.

This protocol is methodologically and conceptually quite different from the three time-reversed protocols described above. In those, participants were required to make quick judgments on each trial with no time to reflect on their decisions. The sequence of events within each trial occurred on a time scale of milliseconds and the putatively causal stimulus appeared immediately after each of the participant's responses. In terms of Kahneman's (2011) dual-mode theory of cognition—as described in his book, *Thinking, Fast and Slow*—these experiments required cognitive processing characteristic of System 1, "Fast Thinking" (also see Evans, 2008, and Evans & Stanovich, 2013).

In contrast, the retroactive facilitation-of-recall protocol confronted participants with a single extended cognitive task that occurred on a time scale of minutes: Presenting the initial list of words took 2-1/2 minutes; the recall test took up to 5 minutes; and the post-test practice exercises took approximately 7 minutes. This allowed participants time to implement deliberate conscious strategies involving working memory, active rehearsal, and verbal categorization, all cognitive processes characteristic of System 2, "Slow Thinking."

Across all his experiments, Bem reported a mean effect size (d) of 0.22, with a Stouffer Z of 6.66, $p = 2.68 \times 10^{-11}$ (Bem *et al.*, 2011).

Bem's experiments have been extensively debated and critiqued. The first published critique appeared in the same issue of the journal as Bem's original article (Wagenmakers *et al.*, 2011). These authors argued that a Bayesian analysis of Bem's results did not support his psi-positive conclusions and recommended that all research psychologists abandon frequentist analyses in favor of Bayesian ones. Bem *et al.* (2011) replied to Wagenmakers *et al.*, criticizing the particular Bayesian analysis they had used and demonstrating that a more reasonable Bayesian analysis yields the same conclusions as Bem's original frequentist analysis. In a similar critique, Rouder & Morey (2011) also advocated a Bayesian approach, criticizing the analyses of both Bem and Wagenmakers *et al.* Rather than continuing to debate this issue in the context of Bem's original experiments, we here analyze the current database with both a frequentist analysis and the specific Bayesian analysis recommended by Rouder and Morey for meta-analyses.

Recently, Judd *et al.* (2012) have argued that psychologists should start treating stimuli statistically as a random factor the same way we currently treat participants. As they acknowledge, this would constitute a major change in practice for psychologists. To illustrate, they re-analyzed several published datasets from psychological journals, including one of Bem's retroactive priming results, showing that when stimuli are treated as a random factor the results are statistically weaker than reported in the original articles. They conclude that "As our simulations make clear, in many commonly used designs in social cognitive research, a likely consequence of only treating participants as a random effect is a large inflation of Type I statistical errors, well above the nominal .05 rate (p. 12)."

Francis (2012) and Schimmack (2012) take a different tack. Instead of arguing that Bem's results are weaker than he reports, they argue that, on the contrary, his results are actually too good to be true. That is, given the statistical power of Bem's effects, it is unlikely that eight of his nine experiments would have achieved statistical significance, implying that there is a hidden file-drawer of experiments or failed statistical analyses that Bem failed to report.

In his own discussion of potential file-drawer issues, Bem (2011) reported that they arose most acutely in his two earliest experiments (on retroactive habituation) because they required extensive pre-experiment pilot testing to select and match pairs of photographs and to adjust the number and timing of the repeated subliminal stimulus exposures. Once these were determined, however, the protocol was "frozen" and the formal experiments begun. Results from the first experiment were used to rematch several of the photographs used for its subsequent replication. In turn, these two initial experiments provided data relevant for setting the experimental procedures and parameters used in all the subsequent experiments.

As Bem's explicitly stated in his article, he omitted one exploratory experiment conducted after he had completed the original habituation experiment and its successful replication. It used supraliminal rather than subliminal exposures. He noted that this fundamentally alters the participant's phenomenology of the experiment, transforming the task into an explicit ESP challenge and thereby undermining the very rationale for using an implicit response measure of psi in the first place. Even that experiment was not left languishing in a file drawer, however, because he had reported and critiqued it at a meeting of the Parapsychological Association (Bem, 2003).

With regard to unreported data analyses, Bem analyzed and reported each experiment with two to four different analyses, demonstrating in each case that the results and conclusions were robust across different kinds of analyses, different indices of psi performance, and different definitions of outliers. Following standard practice, however, he did not treat stimuli as a random factor in his analyses.

In his own critique, Francis (2012) remarks that "perhaps the most striking characteristic of [Bem's] study is that [it meets] the current standards of experimental psychology. The implication is that it is the standards and practices of the field that are not operating properly (p. 155)." Similarly, LeBel & Peters (2011) remark that "...[i]t is precisely because Bem's report is of objectively high quality that it is diagnostic of potential problems with MRP [Modal Research Practice].... Bem has put empirical psychologists in a difficult position: forced to consider either revising beliefs about the fundamental nature of time and causality or revising beliefs about the soundness of MRP (p. 371)."

LeBel and Peters conclude by recommending that we should put a stronger emphasis on replication. We agree. Rather than continuing to debate Bem's original experiments, we seek in our meta-analysis to answer the one question that most decisively trumps such disputes: Can independent investigators replicate the original experiments?

Method

The methodology and reporting of results comply with the Meta-Analysis Reporting Standards (APA, 2008). Additional materials needed to replicate our results independently can be found at http://figshare.com/articles/Meta-analysis_Implicit_Behavioral_Anticipation/903716.

Retrieval and coding of experiments

As noted above, the archival summary publication of Bem's experiments appeared in 2011, but he had begun his first experiments as early as 2000, and began reporting results soon thereafter at departmental colloquia and annual meetings of the Parapsychological Association (Bem, 2003; Bem, 2005; Bem, 2008). Simultaneously he made materials available to anyone expressing an interest in trying to replicate the experiments. As a result, attempted replications of the experiments began to appear as early as 2001 (as reported in Moulton & Kosslyn, 2011).

No presentiment experiments are included in our database because, as noted above, a meta-analysis of those has already been published (Mossbridge *et al.*, 2012). We have, however, included 19 attempted replications of Bem's Retroactive-Facilitation-of Recall experiment that had been previously meta-analyzed by Galak *et al.* (2012) because 8 additional replication studies of that protocol have been reported since then. (This was the only protocol included in Galak *et al.*'s. meta-analysis.)

Although the individual-difference variable of "stimulus seeking" emerged as a significant correlate of psi performance in several of Bem's original experiments, we have not analyzed that variable in the present meta-analysis because too few of the replications reported on it—especially those that modified Bem's original protocol.

Co-authors PT, TR, and MD conducted a search for all potentially relevant replications that became available between the year 2000 and September of 2013. These included unpublished reports as well as peer-reviewed, published articles in mainstream psychological journals; specialized journals; proceedings from conferences; and relevant studies found in Google Scholar, PubMed and PsycInfo. The same set of keywords—*Bem, feeling the future, precognition*— was used for all searches, and no MESH terms or

Boolean operators were used. Using email and academia.edu, they also contacted known psi researchers and mainstream researchers who had expressed an interest in replicating Bem's experiments. Of the ninety-three experiments retrieved, two were eliminated because they were severely underpowered: the first had only one participant; the second had nine (Snodgrass, 2011). A third experiment, reporting positive results, rested on several post-hoc analyses, and so we deemed it too exploratory to include in the meta-analysis (Garton, 2010). The final database thus comprises 90 experiments.

Co-authors PT and TR independently coded and categorized each study with respect to the following variables: a) type of effect(s) tested; b) number of participants enrolled in the study; c) descriptive or inferential statistics used to calculate measures of effect size; d) whether the replication had been conducted before or after the January, 2011 (*Online First*) publication of Bem's original experiments; e) whether or not the experiment had been peer-reviewed; and f) type of replication.

For this last variable, each experiment was categorized into one of three categories: an exact replication of one of Bem's experiments (31 experiments), a modified replication (38 experiments), or an independently designed experiment that assessed the ability to anticipate randomly-selected future events in some alternative way (11 experiments). To qualify as an exact replication, the experiment had to use Bem's software without any procedural modifications other than translating on-screen instructions and stimulus words into a language other than English if needed. The eleven experiments that had not been designed to replicate any of Bem's experiments included five retroactive-priming experiments and six retroactive-practice experiments.

Percentages of agreement for each of the coding variables ranged from a minimum of 90% for the statistical data to 100% for the classification into one of the three categories of experiments. Discrepancies in coding were resolved by discussion between PT and TR.

Frequentist analysis

All the main inferential statistics, weighted effect-size point estimations with corresponding 95% Confidence Intervals, and combined *z* values were calculated using the Comprehensive Meta-Analysis software v.2 by Borenstein *et al.* (2005). Effect sizes (Hedges' *g*) and their standard errors were computed from *t* test values and sample sizes. (Hedges' *g*, is similar to the more familiar *d* [Cohen, 1988], but pools studies using n - 1 for each sample instead of n. This provides a better estimate for smaller sample sizes.) When *t* test values were not available, we used the effect sizes reported by the authors or estimated them from the descriptive statistics. When more than one dependent variable was measured, a single effect size was calculated by averaging the effect sizes obtained by the different *t* values.

Heterogeneity within each set of experiments using a particular protocol (e.g., the set of retroactive priming experiments) was assessed using I^2 (Huedo-Medina *et al.*, 2006). It estimates the percent of variance across studies due to differences among the true effect sizes. If all the studies are methodologically identical and the subject samples are very similar, then I^2 will be small (< 25%) and a fixed-effect model analysis is justified; otherwise a random-effects model is used (Borenstein *et al.*, 2009).

A fixed-effect model assumes that all the studies using a particular protocol have the same true effect size and that the observed variance of effect sizes across the studies is due entirely to random error within the studies. The random-effects model allows for the possibility that different studies included in the analysis may have different true effect sizes and that the observed variation reflects both within-study and between-study sampling error.

Bayesian analysis

A model comparison Bayesian analysis of an experiment pits a specified experimental hypothesis (H_1) against the null hypothesis (H_0) by calculating the odds that H_1 rather than H_0 is true—$p(H_1)/p(H_0)$—or the reverse. The analysis assumes that each person comes to the data with a subjective prior value for these odds and then adjusts them on the basis of the data to arrive at his or her posterior odds. A Bayesian analysis can be summarized by a number called the Bayes Factor (BF), which expresses the posterior odds independent of any particular individual's prior odds. For example, a BF of 3 indicates that the observed data favor the experimental hypothesis over the null hypothesis by a ratio of 3:1. The posterior odds for a particular individual can then be calculated by multiplying his or her prior odds by BF. For example, a mildly psi-skeptical individual might initially assign complementary probabilities of .2 and .8 to H_1 and H_0, respectively, yielding prior odds of .25. If BF = 3 then the Bayesian formula indicates that this individual's posterior odds should be .75. If BF were to exceed 4, then the posterior odds $p(H_1)/p(H_0)$ would exceed 1, implying that this individual now favors the experimental hypothesis over the null.

Jeffreys (1998) has suggested the following verbal labels for interpreting BF levels of $p(H_1)/p(H_0)$:

BF = 1 – 3:	Worth no more than a bare mention
BF = 3 – 10:	Substantial evidence for H_1
BF = 10 – 30:	Strong evidence for H_1
BF = 30 – 100:	Very Strong evidence for H_1
BF > 100:	Decisive evidence for H_1

To perform a Bayesian analysis, one must also specify a prior probability distribution of effect sizes across a range for both H_0 and H_1. Specifying the effect size for H_0 is simple because it is a single value of 0, but specifying H_1 requires specifying a probability distribution across a range of what the effect size might be if H_1 were in fact true. This specification can strongly impact the subsequent estimates of BF and, in fact, was the major disputed issue in the debate over Bem's original experiments (Bem *et al.*, 2011; Rouder & Morey, 2011; Wagenmakers *et al.*, 2011).

For purposes of meta-analysis, Rouder & Morey (2011) argue that one should use the Jeffrey, Zellner and Siow (JZS) prior probability distribution (see, also, Bayarri & Garcia-Donato, 2007). That distribution is designed to minimize assumptions about the range of effect sizes and, in this sense, constitutes what is known as an "objective" prior (Rouder *et al.*, 2009). Moreover, the resulting BF is independent of the measurement scale of the dependent variable, is always finite for finite data, and is consistent in the sense that as sample size increases, BF grows to infinity if the null is false and shrinks to zero if it is true—a consistency that does not obtain for p values. Researchers can also incorporate their expectations for different experimental contexts by tuning the scale of the prior on effect size (designated as r). Smaller values of r (e.g., 0.1) are appropriate when small effects sizes are expected; larger values of r (e.g., 1.0) are appropriate when large effect sizes are expected. As r increases, BF provides increasing support for the null.

For these several reasons, we have adopted the JZS prior probability distribution for our Bayesian analysis. For the estimation of Bayes Factors, we used the meta.ttest function of the BayesFactor package (Morey & Rouder, 2014). In the expectation that the effect size will be small, we set $r = 0.1$. To estimate the overall effect size and τ^2, a measure of between-studies variance, we employed the DiMaggio (2013) script, which uses the R2jags package to run the "BUGS" program (Bayesian Analysis Using Gibb's Sampling). This provides a Monte Carlo Markov Chain simulation approach to parameter estimation using a normally distributed prior with a mean of 0.1 and a wide variance of 10^5. The program chooses samples using either Gibbs or Metropolis Hasting algorithms. Because this is a simulation-based approach, we repeated many draws or iterations and evaluated whether the chain of sample values converged to a stable distribution, which was assumed to be the posterior distribution in which we are interested.

We ran two 20,000 Markov Chain Monte Carlo iterations, each starting with different and dispersed initial values for the model. We based our results on the final 20,000 iterations and assessed whether the chain of values had converged to a stable posterior distribution by monitoring and assessing a graph of the chain and by calculating the Brooks Gelman and Rubin statistic, a tool within the CODA package of R programs for this purpose. The results are presented as mean values of the posterior distributions and their 95% credible intervals (CrI).

Results and discussion

The complete database comprises 90 experiments conducted between 2001 and 2013. These originated in 33 different laboratories located in 14 countries and involved 12,406 participants. The full database with corresponding effect sizes, standard errors, and category assignments is presented in Table S1 along with a forest plot of the individual effect sizes and their 95% confidence intervals.

Dataset 1. Table S1

http://dx.doi.org/10.5256/f1000research.7177.d105136

Experiments in the meta-analysis, N, task type, effect size, standard error, peer-review and replication classifications (Tressoldi *et al.*, 2015).

The first question addressed by the meta-analysis is whether the database provides overall evidence for the anomalous anticipation of random future events. As shown in the first and second rows of Table 1, the answer is yes: The overall effect size (Hedges' g) is 0.09, combined $z = 6.33$, $p = 1.2 \times 10^{-10}$. The Bayesian BF value is 5.1×10^9, greatly exceeding the criterion value of 100 that is considered to constitute "decisive evidence" for the experimental hypothesis (Jeffreys, 1998). Moreover, the BF value is robust across a wide range of the scaling factor r, ranging from a high value of 5.1×10^9 when we set $r = 0.1$ to a low value of 2.0×10^9 when $r = 1.0$.

The second question is whether independent investigators can successfully replicate Bem's original experiments. As shown in the third and fourth rows of Table 1, the answer is again yes: When

Table 1. Meta-analytic results for all experiments and for independent replications of Bem's experiments.

	Number of experiments	Number of participants	Effect size (Hedges' g)	95%CI or CrI	Combined z or Bayes factor	p (One-tailed)	I^2	τ^2
All experiments[a] Bayesian analysis	90	12,406	0.09 0.08	[0.06, 0.11] [0.02, 0.15]	$z = 6.33$ BF = 5.1×10^9	1.2×10^{-10}	41.4	.005 .028
Independent replications[b] Bayesian analysis	69	10,082	0.06 0.07	[0.03, 0.09] [0.01, 0.14]	$z = 4.16$ BF = 3,853	1.2×10^{-5}	36.1	.004 .035
Exact replications Modified replications	31 38	2,106 7,976	0.08 0.05	[0.02, 0.13] [0.02, 0.09]	$z = 2.90$ $z = 3.00$	.0018 .0013	31.7 38.9	.007 .004
Pre-2011 replications Post-2011 replications	30 39	2,193 7,889	0.09 0.05	[0.04, 0.15] [0.02, 0.08]	$z = 3.20$ $z = 2.88$	.0007 .004	39.5 32.3	.009 .003
Peer reviewed Not peer reviewed	35 34	7,477 2,605	0.06 0.06	[0.02, 0.10] [0.02, 0.10]	$z = 2.93$ $z = 3.21$	.0017 .0007	51.4 8.7	.001 .006

Note. In a Bayesian analysis, the analogue to the 95%CI is CrI, "credible intervals of the posterior distributions." I^2 is an estimate of the percent of variance across studies due to differences among the true effect sizes. τ^2 is the between-studies variance.

[a] Assuming a null ES of .01 and a variance of .005 (the observed variance, τ^2, in the random-effects model), the statistical power of this meta-analysis is 0.95 (Hedges & Pigott, 2001).

[b] These analyses exclude Bem's own experiments and the eleven experiments that had not been designed as replications of those experiments.

Bem's experiments are excluded, the combined effect size for attempted replications by other investigators is 0.06, z = 4.16, $p = 1.1 \times 10^{-5}$, and the BF value is 3,853, which again greatly exceeds the criterion value of 100 for "decisive evidence."

The fifth and sixth rows of Table 1 show that the mean effect sizes of exact and modified replications are each independently significant and not significantly different from each other (Mean diff = 0.025; 95% CI [-0.04, 0.09]; z = 0.87, *ns*).

The seventh and eighth rows show that the mean effect sizes of replications conducted before and after the January, 2011 (online) publication of Bem's article are each independently significant and not significantly different from each other (Mean diff = 0.042; 95% CI [.02, 0.10]; z = 0.37, *ns*).

And finally, the bottom two rows of Table 1 show that the mean effect sizes of peer reviewed and not-peer-reviewed replications are each independently significant and identical to each other.

Table 2 displays the meta-analysis of the complete database as a function of experiment type and divided post-hoc into fast-thinking and slow-thinking protocols.

As shown in Table 2, fast-thinking protocols fared better than slow-thinking protocols: Every fast-thinking protocol individually achieved a statistically significant effect, with an overall effect size of 0.11 and a combined z greater than 7 sigma. In contrast, slow-thinking experiments achieved an overall effect size of only 0.03, failing even to achieve a conventional level of statistical significance (p = .16).

One possible reason for the less successful performance of the slow-thinking experiments is that 12 of the 27 attempted replications of Bem's retroactive facilitation of recall experiment were modified replications. The 15 exact replications of that protocol yielded an overall effect size of 0.08, but the 12 modified replications yielded a null effect size (-0.00). For example, Galak *et al.* (2012) used their own software to conduct seven of their 11 modified replications in which 87% of the sessions (2,845 of 3,289 sessions) were conducted online, thereby bypassing the controlled conditions of the laboratory. These unsupervised sessions produced an overall effect size of -0.02. Because experiments in a meta-analysis are weighted by sample size, the huge N of these online experiments substantially lowers the mean effect size of the replications: When the online experiments are removed, the mean ES for this protocol rises to 0.06 [0.00, 0.12]; z = 1.95, p = .05.

Nevertheless, we still believe that it is the fast/slow variable itself that is an important determinant of the lower success rate of the slow-thinking experiments. In particular, we suspect that fast-thinking protocols are more likely to produce evidence for psi because they prevent conscious cognitive strategies from interfering with the automatic, unconscious, and implicit nature of psi functioning (Carpenter, 2012). This parallels the finding in conventional psychology that mere exposure effects are most likely to occur when the exposures are subliminal or incidental because the participant is not aware of them and, hence, is not prompted to counter their attitude-inducing effects (Bornstein, 1989).

Finally, Table 2 reveals that the clear winner of our meta-analytic sweepstakes is the precognitive detection of erotic stimuli (row 1), the time-reversed version of psychology's time-honored Law of

Table 2. Meta-analytic results as a function of protocol and experiment type.

Experiment Type	Number of experiments	Number of participants	Effect size	95%CI	Combined z	p (One-tailed)	I^2
Fast-thinking protocols							
Precognitive detection of reinforcement	14	863	0.14[a]	[0.08, 0.21]	4.22	1.2×10^{-5}	19.0
Precognitive avoidance of negative stimuli	8	3,120	0.09	[0.03, 0.14]	3.10	.002	50.5
Retroactive priming	15	1,154	0.11	[0.03, 0.21]	2.85	.003	42.0
Retroactive habituation	20	1,780	0.08[a]	[0.04, 0.13]	3.50	.0002	24.6
Retroactive practice	4	780	0.11[a]	[0.04, 0.18]	3.03	.002	00.0
All fast-thinking experiments	**61**	**7,697**	**0.11**	**[0.08, 0.14]**	**7.11**	**5.8×10^{-13}**	**31.6**
Slow-thinking protocols							
Retroactive facilitation of practice on recall	27	4,601	0.04	[-0.01, 0.09]	1.66	.10	38.3
Retroactive facilitation of practice on text reading speed	2	108	-0.10	[-0.40, 0.20]	-0.65	.51	61.0
All slow-thinking experiments	**29**	**4,709**	**0.03**	**[-0.01, 0.08]**	**1.38**	**.16**	**39.7**

[a] Fixed-effect model

Effect. The fourteen experiments using that protocol— conducted in laboratories in four different countries—achieve a larger effect size (0.14), a larger combined *z* (4.22), and a more statistically significant result ($p = 1.2 \times 10^{-5}$) than any other protocol in the Table. This protocol was also the most reliable: If we exclude the three experiments that were not designed to be replications of Bem's original protocol, 10 of the 11 replication attempts were successful, achieving effect sizes ranging from 0.12 to 0.52. The one exception was a replication failure conducted by Wagenmakers *et al.* (2012), which yielded a non-significant effect in the unpredicted direction, ES = -0.02, *t*(99) = -0.22, *ns*. These investigators wrote their own version of the software and used a set of erotic photographs that were much less sexually explicit than those used in Bem's experiment and its exact replications.

The results of our meta-analysis do not stand alone. As we noted in the introduction, Bem's experiments can be viewed as conceptual replications of the presentiment experiments in which participants display physiological arousal to erotic and negative photographs a few seconds before the photographs are selected and displayed (Mossbridge *et al.*, 2012). The parallel is particularly close for the two protocols testing the precognitive detection of erotic stimuli and the precognitive avoidance of negative stimuli (Protocols 1 and 2 in Table 2). Together those two protocols achieve a combined effect size of 0.11, $z = 4.74$, $p = 1.07 \times 10^{-6}$.

File-drawer effects: Selection bias and *P*-hacking

Because successful studies are more likely to be published than unsuccessful studies—the file-drawer effect—conclusions that are drawn from meta-analyses of the known studies can be misleading. To help mitigate this problem, the Parapsychological Association adopted the policy in 1976 of explicitly encouraging the submission and publication of psi experiments regardless of their statistical outcomes. Similarly, we put as much effort as we could in locating unpublished attempts to replicate Bem's experiments by contacting both psi and mainstream researchers who had requested his replication packages or had otherwise expressed an interest in replicating the experiments. As we saw in Table 1, this all appears to have had the desired effect on the current database: Peer-reviewed experiments yielded the same results as experiments that were not peer-reviewed.

There are also several statistical techniques for assessing the extent to which the absence of unknown studies might be biasing a meta-analysis. We consider nine of them here.

Fail-safe calculations

One of the earliest of these techniques was the calculation of a "Fail-Safe *N*," the number of unknown studies averaging null results that would nullify the overall significance level of the database if they were to be included in the meta-analysis (Rosenthal, 1979). The argument was that if this number were implausibly large, it would give us greater confidence in the conclusions based on the known studies. The Rosenthal Fail-Safe *N*, however, has been criticized as insufficiently conservative because it does not take into account the likely possibility that unpublished or unretrieved studies might well have a mean non-zero effect in the unpredicted direction. Thus the estimate of the Fail-Safe *N* is likely to be too high. (For the record, the Rosenthal Fail-Safe *N* for our database is greater than 1,000.)

An alternative approach for estimating a Fail-Safe *N* focuses on the effect size rather than the *p* value (Orwin, 1983). The investigator first specifies two numbers: The first is an average effect size for missing studies which, if added to the database, would bring the combined effect size under a specified "trivial" threshold—the second number that must be specified. If we set the mean effect size of missing studies at .001 and define the threshold for a "trivial" effect size to be .01, then the Orwin Fail-Safe *N* for our database is 544 studies. That is, there would have to be 544 studies missing from our database with a mean effect size of .001 to reduce its overall effect size to .01.

Correlations between study size and effect size

Another set of indices for assessing selection bias are various correlational measures for assessing the relationship between the size of a study and its effect size. The most direct is the Begg and Mazumdar's rank correlation test, which simply calculates the rank correlation (Kendall's tau) between the variances or standard errors of the studies and their standardized effect sizes (Rothstein *et al.*, 2005). If this correlation is significantly negative, if small underpowered studies have larger effect sizes than larger studies, then there is reason to suspect the presence of publication or retrieval bias in the database. For our database, Kendall's tau is actually slightly positive: $\tau = +0.10$; $z = 1.40$, implying that our database is not seriously biased by a selection bias.

More recent publications (e.g., Jin *et al.*, 2015; Rücker *et al.*, 2011; Schwarzer *et al.*, 2010; Stanley & Doucouliagos, 2014; Stanley & Doucouliagos, 2015) have urged the adoption of more complex indices of selection bias:

1. The Copas method (Copas, 2013; Schwarzer *et al.*, 2010) is based on two models, the standard random effects model and the selection model, which takes study size into account.

2. The Limit meta-analysis (Schwarzer *et al.*, 2014) is an extended random effects model that takes account of possible small-study effects by allowing the treatment effect to depend on the standard error.

3. The Precision Effect Test (PET, Stanley, 2008; Stanley & Doucouliagos, 2014) is a variant of the classical Egger regression test (Sterne & Egger, 2005), which tests the relationship between study size and effect size.

4. The Weighted Least Squares analysis (Stanley & Doucouliagos, 2015) provides estimates that are comparable to random effects analyses when there is no publication bias and are identical to fixed-effect analyses when there is no heterogeneity, providing superior estimates compared with both conventional fixed and random effects analyses.

Table 3 summarizes the results of applying these four additional tests to our database.

As Table 3 shows, three of the four tests yield significant effect sizes estimates for our database after being corrected for potential selection bias; the PET analysis is the only test in which the 95% confidence interval includes the zero effect size. As Sterne & Egger (2005) themselves caution, however, this procedure cannot assign a causal mechanism, such as selection bias, to the correlation between study size and effect size, and they urge the use of the more noncommittal term "small-study effect."

Trim and fill

Currently the most common method for estimating the number of studies with low effect sizes that might be missing from a database is Duval & Tweedie's (2000) Trim-and-Fill procedure. It is based on a graphic display of the correlation between sample size and effect size called the "funnel" plot, which plots a measure of sample size on the vertical axis as a function of effect sizes on the horizontal axis. The funnel plot for our database is displayed in Figure 1, which uses the reciprocal of the standard error as the measure of sample size.

If a meta-analysis has captured all the relevant experiments, we would expect the funnel plot to be symmetric: Experiments should be dispersed equally on both sides of the mean effect size. If the funnel plot is asymmetric, with a relatively high number of small experiments falling to the right of the mean effect size and relatively few falling to the left, it signals the possibility that there may be experiments with small or null effects that actually exist but are missing from the database under consideration.

Using an iterative procedure, the trim-and-fill method begins by trimming experiments from the extreme right end of the plot (i.e., the smallest studies with the largest effect sizes) and then calculating a new mean effect size. It then reinserts the trimmed studies on

Table 3. Copas method, Limit meta-analysis, Precision Effect Test and Weighted least squares results for the overall and the "fast-thinking" database.

Test		Effect size estimate	95%CI
Copas method	Overall	0.08	[0.05, 0.10]
	Fast-thinking	0.07	[0.03, 0.10]
Limit meta-analysis	Overall	0.05	[0.02, 0.08]
	Fast-thinking	0.05	[0.01, 0.10]
Precision Effect Test (PET)	Overall	0.01	[-0.04, 0.05]
	Fast-thinking	0.03	[-0.03, 0.08]
Weighted Least Squares	Overall	0.06	[0.04, 0.09]
	Fast-thinking	0.09	[0.06, 0.12]

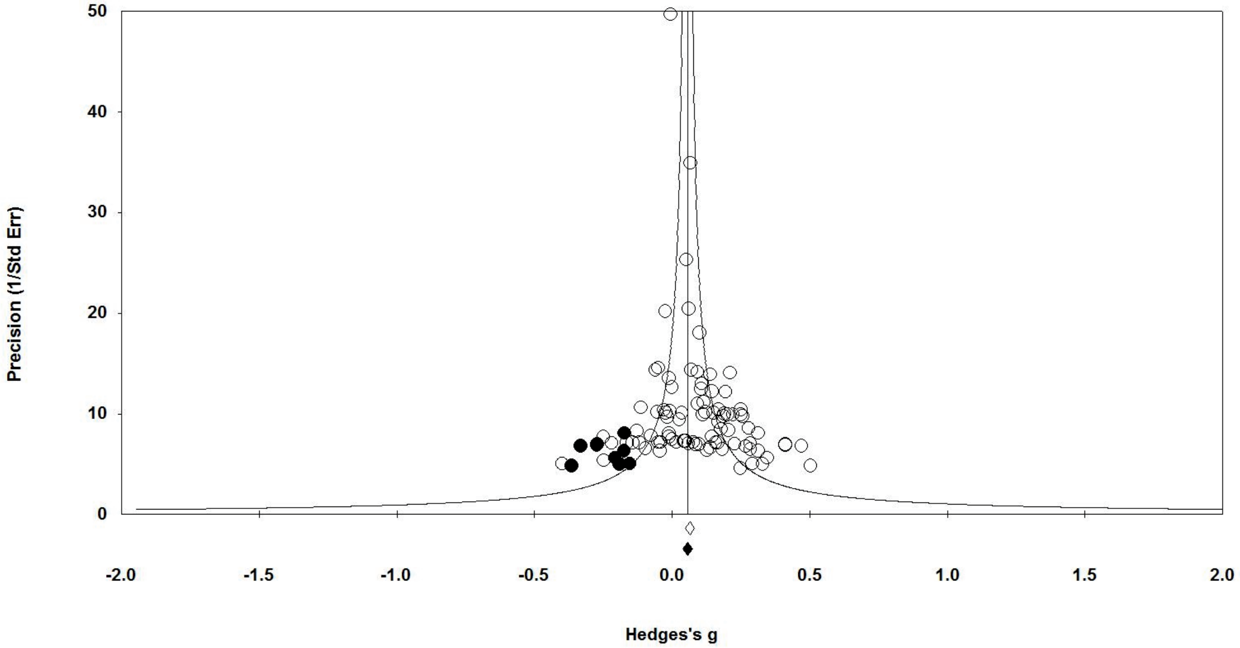

Figure 1. Funnel Plot of the observed studies (white circles) and the imputed missing studies (black circles) under a random-effects model.

the right and inserts their imputed "missing" counterparts symmetrically to the left of the new mean effect size. This produces a revised, more symmetric funnel plot centered around the newly revised mean effect size. This process continues until the funnel plot becomes symmetric. At that point, the plot is centered around a final corrected estimate of the effect size and displays the number of imputed "missing" experiments to the left of the unbiased mean effect size.

Figure 1 displays the funnel plot for our complete database after it has been modified by the trim-and-fill procedure. The unfilled diamond under the horizontal axis marks the original observed effect size (0.09, see Table 1) and the black diamond marks the corrected estimate of the effect size: 0.07 [0.04, 0.10]. The unfilled circles identify the 90 actual experiments in the meta-analysis; the black circles identify the imputed missing experiments. As Figure 1 shows, there are only eight potentially missing studies. As noted above, the Orwin Fail-Safe estimate of how many missing experiments with low effect sizes would be required to nullify the overall effect size of the database is 544.

P-curve analysis

All the analyses discussed above presume that selection bias is driven by effect-size considerations, but Simonsohn *et al.* (2014a); Simonsohn *et al.* (2014b) have argued that it is actually more likely to be driven by the $p = .05$ significance level. They have also demonstrated empirically that the trim and fill procedure is inadequate for estimating the true effect size present in the database (2014b). In its place, they and other authors (van Assen *et al.*, 2015) have recently proposed a very different approach called *p*-curve analysis.

P-curve is the distribution of significant ($p < .05$) results among the experiments in a meta-analysis. "It capitalizes on the fact that the distribution of significant *p* values... is a function of the true underlying effect. Researchers armed only with sample sizes and test results of the published findings can correct for publication bias (Simonsohn *et al.*, 2014b, p. 666)." In addition to assessing selection bias, *p*-curve analysis can also assess the presence of "*p*-hacking," questionable practices of selective reporting that illegitimately enable an investigator to claim results that meet the coveted $p < .05$ threshold (Simonsohn, *et al.*, 2014a; Simonsohn, *et al.*, 2014b).

In our database, 17 (19%) of the 90 studies reported results that were statistically significant at the .05 level. The solid blue line in Figure 2 displays the *p*-curve distribution of those studies.

The dotted horizontal red line ("Null of zero effect") is the distribution expected if there is no effect in the data. In that case, 5% of the significant *p* values will be below .05, 4% will be below .04, 3% will be below .03, 2% will be below .02, and 1% will be below .01. Thus there will be as many *p* values between .04 and .05 as between .00 and .01, and the shape of the *p*-curve is a uniform, straight horizontal line with 20% of the significant values within each of the 5 intervals on the horizontal-axis. If a genuine non-zero effect exists, however, then *p*-curve's expected distribution will be right-skewed:

> We expect to observe more low significant *p* values ($p < .01$) than high significant *p* values ($.04 < p < .05$) (Simonsohn *et al.*, 2014b, pp. 666–667)... A set of significant findings contains evidential value when we can rule out selective reporting as the sole explanation of those findings. Only right-skewed *p*-curves... are diagnostic of evidential value. *P*-curves that are not right-skewed suggest that the set of findings lacks evidential value, and curves that are left-skewed suggest the presence of intense *p*-hacking (Simonsohn *et al.*, 2014a, p. 535).

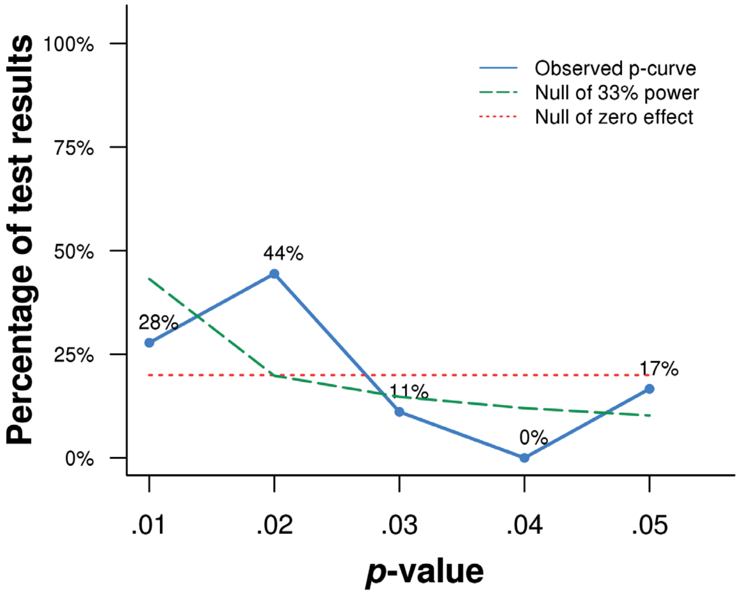

Figure 2. Distribution of the significant *p* values across experiments in the meta-analysis.

Table 4 presents the skewness analysis of our database (Simonsohn *et al.*, 2014b).

As shown in the first row of Table 4, the right-skew of the *p*-curve is equivocally significant (*p* = .048, *p* = .056). When this is the case, Simonsohn *et al.* (2014a) propose applying a second test to see if the studies lack evidential value because they are flatter than an underpowered (33%) *p*-curve—depicted by the dashed green line. As shown in the second row of the Table, the observed *p*-curve is *not* flatter than the null at 33% power, so we cannot conclude that the evidential value of the database is inadequate. And finally, the bottom row shows that the *p*-curve is clearly not left-skewed, implying that the database has not been strongly *p*-hacked.

Because the right-skew of the *p*-curve is equivocally significant, we turned to a more direct *p*-curve algorithm called *p*-uniform (Van Assen *et al.*, 2015) that directly tests the degree to which the observed curve differs from the "no-effect" uniform distribution. (If there is a substantial amount of heterogeneity in the meta-analysis, this method should be used as a sensitivity analysis.) The *p*-uniform test confirms that there is, in fact, a significant effect in our database (*p* = .005) and that there is no evidence for selection bias (*p* = .857).

In sum, eight of the nine statistical tests we have applied to our database support the conclusion that its overall statistical significance has not been compromised by either selection bias or by *p*-hacking.

P-curve and the true effect size

One of the counterintuitive derivations from *p*-curve analysis—confirmed by extensive simulations—is that when the distribution of significant *p* values is right-skewed, the inclusion of studies with nonsignificant *p* levels (*p* > .05) in a meta-analysis actually underestimates the true effect size in the database (Simonsohn *et al.*, 2014b). Based on the Simonsohn *et al.* *p*-curve analysis, the estimate of the true effect size for our database is 0.20, virtually identical to the mean effect size of Bem's (2011) original experiments (0.22) and the mean effect size of the presentiment experiments (0.21) (Mossbridge *et al.*, 2012). A comparable calculation cannot be legitimately derived from the *p*-uniform algorithm because it assumes that the population effect size is fixed rather than heterogeneous (van Assen *et al.*, 2015, p. 4). As shown in Table 1, our population effect size is heterogeneous.

The complementary merits of exact and modified replications

Our meta-analysis reveals that both exact and modified replications of Bem's experiments achieve significant and comparable success rates (Table 1). This is reassuring because the two kinds of replication have different advantages and disadvantages. When a replication succeeds, it logically implies that every step in the replication "worked." When a replication fails, it logically implies that at least one or more of the steps in the replication failed—including the possibility that the experimental hypothesis is false—but we do not know which step(s) failed. As a consequence, even when exact replications fail, they are still more informative than modified replications because they dramatically limit the number of potential variables that might have caused the failure.

There is, of course, no such thing as a truly exact replication. For example, the experimenter's attitudes and expectations remain uncontrolled even in a procedurally exact replication, and there are now more than 345 experiments demonstrating that experimenter attitudes and expectations can produce belief-confirming results, even in simple maze experiments with rats as subjects (Rosenthal & Rubin, 1978).

Exact replications also serve to guard against some of the questionable research practices that can produce false-positive results, such as changing the protocol or experimental parameters as the experiment progresses, selectively reporting comparisons and covariates without correcting for the number examined, and selectively presenting statistical analyses that yielded significant results while omitting other analyses that did not (Simmons *et al.*, 2011). By defining an exact replication in our meta-analysis as one that used Bem's experimental instructions, software, and stimuli, we ensure that the experimental parameters and data analyses are all specified ahead of time. In other words, an exact replication is a publicly available, pre-specified protocol that provides many of the same safeguards against false-positive results that are provided by the preregistration of planned experiments.

Table 4. Skewness tests on the distribution of significant *p* values across experiments in the database.

Statistical Inference	Binomial Test	Continuous Test
Studies contain evidential value. (*Right-skew*)	*p* = .048	z = -1.58, *p* = .056
Studies' evidential value, if any, is inadequate. (*Flatter than 33% power*)	*p* = .64	z = -0.79, *p* = .16
Studies exhibit evidence of intense *p*-hacking. (*Left skew*)	*p* = .98	z = 1.58, *p* = .94

Despite the merits of exact replications, however, they cannot uncover artifacts in the original protocol that may produce false positive results, whereas suitably modified replications can do exactly that by showing that an experiment fails when a suspected artifact is controlled for. Modified replications can also assess the generality of an experimental effect by changing some of the parameters and observing whether or not the original results are replicated. For example, the one failed replication of the erotic stimulus detection experiment (Wagenmakers *et al.*, 2012) had substituted mild, non-explicit erotic photographs for the more explicit photographs used in Bem's original experiment and its exact replications.

As we noted in the introduction, Judd *et al.* (2012) have recently suggested that psychologists should begin to treat stimuli statistically as a random factor the same way we currently treat participants. This would constitute a way of testing the generalizability of results in psychological experiments. This would, however, also represent a major change in current practice in psychology, and none of the experiments in our database treated stimuli as a random factor. Nevertheless, some generality of stimuli used in Bem's experimental protocols is achieved. In those involving erotic photographs, for example, different stimulus sets are used for men and women and all participants are given the choice of viewing opposite-sex or same-sex erotica. Experiments using words as stimuli (e.g., retroactive priming experiments) were successfully replicated in languages other than English.

The fact that exact and modified replications of Bem's experiments produced comparable, statistically significant results thus implies generality across stimuli, protocols, subject samples, and national cultures. Moreover, the different protocols can themselves be viewed as conceptual replications of the overarching hypothesis that individuals are capable of anomalously anticipating random future events.

General discussion

As Bem noted in his original 2011 article, psi is a controversial subject, and most academic psychologists do not believe that psi phenomena are likely to exist. A survey of 1,188 college professors in the United States revealed that psychologists were much more skeptical about psi than respondents in the humanities, the social sciences, or the physical sciences, including physics (Wagner & Monnet, 1979). Although this survey is now several years old, many psi researchers have observed that psychologists continue to be the most psi-skeptical subgroup of academics.

As Bem further noted, there are, in fact, justifiable reasons for the greater skepticism of psychologists. Although our colleagues in other disciplines would probably agree with the oft-quoted dictum that "extraordinary claims require extraordinary evidence," we psychologists are more likely to be familiar with the methodological and statistical requirements for sustaining such claims and aware of previous claims that failed either to meet those requirements or to survive the test of successful replication. Even for ordinary claims, our conventional frequentist statistical criteria are conservative: The $p = .05$ threshold is a constant reminder that it is worse to assert that an effect exists when it does not (the Type I error) than to assert that an effect does not exist when it does (the Type II error). (For a refreshing challenge to this view, see Fiedler *et al.*, 2012).

Second, research in cognitive and social psychology over the past 40 years has sensitized us psychologists to the errors and biases that plague intuitive attempts to draw valid inferences from the data of everyday experience (e.g. Gilovich, 1991; Kahneman, 2011). This leads us to give virtually no weight to anecdotal or journalistic reports of psi, the main source cited in the survey by our colleagues in other disciplines as evidence for their more favorable beliefs about psi.

One sobering statistic from the survey was that 34% of psychologists in the sample asserted psi to be impossible, more than twice the percentage of all other respondents (16%). Critics of Bayesian analyses frequently point out the *reductio ad absurdum* case of the extreme skeptic who declares psi or any other testable phenomenon to be impossible. The Bayesian formula implies that for such a person, no finite amount of data can raise the posterior probability in favor of the experimental hypothesis above 0, thereby conferring illusory legitimacy on the most anti-scientific stance. More realistically, all an extreme skeptic needs to do is to set his or her prior odds in favor of the psi alternative sufficiently low so as to rule out the probative force of any data that could reasonably be proffered.

Which raises the following question: On purely statistical grounds, are the results of our meta-analysis strong enough to raise the posterior odds of such a skeptic to the point at which the psi hypothesis is actually favored over the null, however slightly?

An opportunity to calculate an approximate answer to this question emerges from the Bayesian critique of Bem's original experiments made by Wagenmakers *et al.* (2011). Although they did not explicitly claim psi to be impossible, they came very close by setting their prior odds at 10^{20} against the psi hypothesis. As shown in Table 1, the Bayes Factor for our database is approximately 10^{9} *in favor* of the psi hypothesis, which implies that our meta-analysis should lower their posterior odds against the psi hypothesis to 10^{11}. In other words, our "decisive evidence" falls 11 orders of magnitude short of convincing Wagenmakers *et al.* to reject the null. (See a related analysis of their prior odds in Bem *et al.*, 2011.) Clearly psi-proponents have their work cut out for them.

Beyond this Bayesian argument, a more general reason that many psychologists may find a meta-analysis insufficiently persuasive is that the methodology of meta-analysis is itself currently under intense re-examination, with new procedural safeguards (e.g. preregistration of all included studies) and statistical procedures (e.g., treating stimuli as a random factor, *p*-curve analysis) appearing almost monthly in the professional literature. Even though our meta-analysis was conceived and initiated prior to many of these developments, we were able to make use of many of them after the fact, (e.g., *p*-curve analysis) but not others (e.g., preregistration, stimuli treated as a random factor). We thus hope that other researchers will be motivated to follow up with additional experiments and analyses to confirm, disconfirm, or clarify the nature of our findings.

Perhaps the most reasonable and frequently cited argument for being skeptical about psi is that there is no explanatory theory or proposed mechanism for psi phenomena that is compatible with current physical and biological principles. Indeed, this limitation is implied by the very description of psi as "anomalous," and it provides an arguably legitimate rationale for imposing the requirement that the evidence for psi be "extraordinary."

We would argue, however, that this is still not a legitimate rationale for rejecting proffered evidence *a priori*. Historically, the discovery and scientific exploration of most phenomena have preceded explanatory theories, often by decades (e.g., the analgesic effect of aspirin; the anti-depressant effect of electroconvulsive therapy) or even centuries (e.g., electricity and magnetism, explored in ancient Greece as early as 600 BC, remained without theoretical explanation until the Nineteenth Century). The incompatibility of psi with our current conceptual model of physical reality may say less about psi than about the conceptual model of physical reality that most non-physicists, including psychologists, still take for granted—but which physicists no longer do.

As is widely known, the conceptual model of physical reality changed dramatically for physicists during the 20th Century, when quantum theory predicted and experiments confirmed the existence of several phenomena that are themselves incompatible with our everyday Newtonian conception of physical reality. Some psi researchers see sufficiently compelling parallels between certain quantum phenomena (e.g., quantum entanglement) and characteristics of psi to warrant considering them as potential mechanisms for psi phenomena (e.g., Broderick, 2007; Radin, 2006). Moreover, specific mechanisms have been proposed that seek to explain psi effects with theories more testable and falsifiable than simple metaphor (e.g., Bierman, 2010; Maier & Buechner, 2015; Walach *et al.*, 2014). A recent collection of these theories is presented in May & Marwaha (2015).

Although very few physicists are likely to be interested in pursuing explanations for psi, the American Association for the Advancement of Science (AAAS) has now sponsored two conferences of physicists and psi researchers specifically organized to discuss the extent to which precognition and retrocausation can be reconciled with current or modified versions of quantum theory. The proceedings have been published by the American Institute of Physics (Sheehan, 2006; Sheehan, 2011). A central starting point for the discussions has been the consensus that the fundamental laws of both classical and quantum physics are time symmetric:

> They formally and equally admit time-forward and time-reversed solutions.... Thus, though we began simply desiring to predict the future from the present, we find that the best models do not require—in fact, do not respect—this asymmetry.... [Accordingly,] it seems untenable to assert that time-reverse causation (retrocausation) cannot occur, even though it temporarily runs counter to the macroscopic arrow of time (Sheehan, 2006, p. vii).

Ironically, even if quantum-based theories of psi eventually do mature from metaphor to genuinely predictive models, they are still not likely to provide intuitively satisfying descriptive mechanisms for psi because quantum theory itself fails to provide such mechanisms for physical reality. Physicists have learned to live with that conundrum in several ways. Perhaps the most common is simply to ignore it and attend only to the mathematics and empirical findings of the theory—derisively called the "Shut Up and Calculate" school of quantum physics (Kaiser, 2012).

As physicist and Nobel Laureate Richard Feynman (1994) advised, "Do not keep saying to yourself... 'but how can it be like that?' because you will get...into a blind alley from which nobody has yet escaped. Nobody knows how it can be like that (p. 123)."

Meanwhile the data increasingly compel the conclusion that it really *is* like that.

Perhaps in the future, we will be able to make the same statement about psi.

Data availability

F1000Research: Dataset 1. Table S1, 10.5256/f1000research.7177.d105136

Author contributions

P. Tressoldi initiated this project, performed the meta-analyses, and shared with D. Bem the primary responsibility for writing this article; T. Rabeyron and M. Duggan were primarily responsible for retrieving and classifying the studies and were also active contributors to the writing of the article. All authors have seen and agreed to the final content of the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

We are grateful to Charles DiMaggio for his collaboration in implementing Bayesian parameter estimation, to Robbie van Aert and Marcel van Assen for performing their *p*-uniform analysis on our database, and to Daniel Lakens for his collaboration in preparing the *p*-curve analysis and critical examination of effect sizes. These helpful colleagues do not necessarily concur with our conclusions.

Supplementary materials

Figure S1.

Forest plot of effect sizes. Each blue dot identifies the estimated effect size for that experiment with the corresponding 95% confidence interval. The red vertical line marks the overall effect size based on the random-effects model.

Click here to access the data.

References

(References marked with a single asterisk indicate studies included in the meta-analysis)

APA Publication and Communication Board Working Group on Journal Article Reporting Standards. **Reporting standards for research in psychology: why do we need them? What might they be?** *Am Psychol.* 2008; **63**(9): 839–851.
PubMed Abstract | Publisher Full Text | Free Full Text

*Barušs I, Rabier V: **Failure to replicate retrocausal recall.** *Psychology of Consiousness: Theory, Research, and Practice.* 2014; **1**(1): 82–91.
Publisher Full Text

*Batthyany A: **A replication of Bem's retro-priming study.** *Personal communication.* 2008.

*Batthyany A: **Retroactive/Precognitive Priming: The role of attention allocation on time-reversed affective processing.** *Personal communication.* 2009.
Reference Source

*Batthyany A: **Retrocausal Habituation and Induction of Boredom: A Successful Replication of Bem (2010; Studies 5 and 7)**. 2010.
Reference Source

*Batthyany A, Kranza GS, Erber A: **Moderating factors in precognitive habituation: the roles of situational vigilance, emotional reactivity, and affect regulation.** *J Soc Psych Res.* 2009; **73**(895[2]): 65–82.
Reference Source

*Batthyany A, Spajic I: **The Time-Reversed Mere Exposure Effect: Evidence for Long-Delay, but not Short-Delay Retrocausal Affective Processing.** *Personal communication.* 2008.

Bayarri MJ, Garcia-Donato G: **Extending conventional priors for testing general hypotheses in linear models.** *Biometrika.* 2007; **94**(1): 135–152.
Publisher Full Text

Bem DJ: **Precognitive habituation; Replicable evidence for a process of anomalous cognition**. *Paper presented at the Parapsychology Association 46th Annual Convention*, Vancouver, Canada. 2003, August 2–4.
Reference Source

Bem DJ: **Precognitive aversion**. *Paper presented at the Parapsychology Association 48th Annual Convention*, Petaluma, CA, 2005, August 11–15.

Bem DJ: **Feeling the future III: Additional experimental evidence for apparent retroactive influences on cognition and affect**. *Paper presented at the Parapsychology Association 51st Annual Convention*, Winchester, England, 2008, August 13–17.

*Bem DJ: **Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect.** *J Pers Soc Psychol.* 2011; **100**(3): 407–425.
PubMed Abstract | Publisher Full Text

*Bem DJ: **An additional replication of the "precognitive detection of erotic stimuli" experiment.** *Personal communication.* 2012.

Bem DJ, Utts J, Johnson WO: **Must psychologists change the way they analyze their data?** *J Pers Soc Psychol.* 2011; **101**(4): 716–719.
PubMed Abstract | Publisher Full Text

*Bierman D: **Anomalous Switching of the Bi-Stable Percept of a Necker Cube: A Preliminary Study.** *J Sci Explor.* 2011; **25**(4): 721–733.
Reference Source

Bierman DJ: **Consciousness induced restoration of time symmetry (CIRTS): A psychophysical theoretical perspective.** *J Parapsychol.* 2010; **24**: 273–300.
Reference Source

Bierman DJ, Radin DI: **Anomalous anticipatory response on randomized future conditions.** *Percept Mot Skills.* 1997; **84**(2): 689–690.
PubMed Abstract | Publisher Full Text

*Bijl A, Bierman D: **Retro-active training of rational vs. intuitive thinkers**. *Paper presented at the 56th Parapsychological Convention*, Viterbo, Italy, 2013.
Reference Source

*Boer De R, Bierman D: **The roots of paranormal belief: divergent associations or real paranormal experiences**? *Proceedings of Presented Papers: The Parapsychological Association 49th Annual Convention*, 2006; 283–298.
Reference Source

Borenstein M, Hedges LV, Higgins JPT, *et al.*: **Introduction to meta-analysis**. Wiley: Chichester, 2009.
Publisher Full Text

Borenstein M, Hedges L, Higgins J, *et al.*: **Comprehensive meta-analysis (Version 2)**. Englewood, NJ: Biostat, 2005.
Reference Source

Bornstein RF: **Exposure and affect: Overview and meta-analysis of research, 1968–1987.** *Psychol Bull.* 1989; **106**(2): 265–289.
Publisher Full Text

Broderick D: **Outside the Gates of Science: Why It's Time for the Paranormal to Come in From the Cold**. New York: Thundermouth Press, 2007.
Reference Source

*Cardeña E, Marcusson-Clavertz D, Wasmuth J: **Hypnotizability and dissociation as predictors of performance in a precognition task: A pilot study.** *J Parapsychol.* 2009; **73**(1): 137–158.
Reference Source

Carpenter JC: **First sight: ESP and parapsychology in everyday life**. Lanham MD: Rowman & Littlefield, 2012.
Reference Source

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

Copas JB: **A likelihood-based sensitivity analysis for publication bias in meta-analysis.** *J R Stat Soc Ser C Appl Stat.* 2013; **62**(1): 47–66.
Publisher Full Text

Dijksterhuis A, Smith PK: **Affective habituation: subliminal exposure to extreme stimuli decreases their extremity.** *Emotion.* 2002; **2**(3): 203–214.
PubMed Abstract | Publisher Full Text

DiMaggio C: **Bayesian Analysis for Epidemiologists Part IV: Meta-Analysis**. Injury Control and Epidemiology Pages at Columbia (ICEPaC). Columbia University, 2013.
Reference Source

Duval S, Tweedie R: **Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis.** *Biometrics.* 2000; **56**(2): 455–463.
PubMed Abstract | Publisher Full Text

Evans JSB: **Dual-processing accounts of reasoning, judgment, and social cognition.** *Annu Rev Psychol.* 2008; **59**: 255–278.
PubMed Abstract | Publisher Full Text

Evans JSB, Stanovich KE: **Dual-Process Theories of Higher Cognition: Advancing the Debate.** *Perspect Psychol Sci.* 2013; **8**(3): 223–241.
PubMed Abstract | Publisher Full Text

Ferguson MJ, Zayas V: **Automatic evaluation.** *Curr Dir Psychol Sci.* 2009; **18**(6): 362–366.
Publisher Full Text

Feynman R: **The character of physical law**. New York, NY: Modern Library, 1994.
Reference Source

Fiedler K, Kutzner F, Krueger JI: **The Long Way From α-Error Control to Validity Proper: Problems With a Short-Sighted False-Positive Debate.** *Perspect Psychol Sci.* 2012; **7**(6): 661–669.
PubMed Abstract | Publisher Full Text

*Fontana G, Polikarpov V, Yankelevich A: **Experiments on anomalous retroactive influences in the context of the theory of growing block universe**. 2012.
Reference Source

Francis G: **Too good to be true: publication bias in two prominent studies from experimental psychology.** *Psychon Bull Rev.* 2012; **19**(2): 151–156.
PubMed Abstract | Publisher Full Text

*Franklin MS, Schooler JW: **Can practice effects extend backwards in time? An overview of 7 years of experimentation**. *Presentation at the 32nd Annual Meeting of the Society for Scientific Exploration*, Dearborn, Michigan, 2013.

*Galak J, Leboeuf RA, Nelson LD, *et al.*: **Correcting the past: failures to replicate ψ.** *J Pers Soc Psychol.* 2012; **103**(6): 933–948.
PubMed Abstract | Publisher Full Text

Garton R: **Precognitive priming and sequential effects in visual word recognition**. Master Thesis, Macquarie University, Australia, 2010.
Reference Source

Gilovich T: **How we know what isn't so: the fallibility of human reason in everyday life**. New York, NY: The Free Press, 1991.
Reference Source

*Hadlaczky G, Westerlund J: **Precognitive Habituation: An Attempt to Replicate Previous Results**. *Paper presented at The 29th International Conference of the Society for Psychical Research*, University of Bath UK, 2005.

Hedges LV, Pigott TD: **The power of statistical tests in meta-analysis.** *Psychol Methods.* 2001; **6**(3): 203–217.
PubMed Abstract | Publisher Full Text

*Hitchman GM, Roe CA, Sherwood SJ: **A re-examination of non-intentional precognition with openness to experience, creativity, psi beliefs and luck beliefs as predictors of success.** *J Parapsychol.* 2012a; **76**(1): 109–145.
Reference Source

*Hitchman GM, Roe CA, Sherwood SJ: **The influence of latent inhibition on performance at a non-intentional precognition task**. *Proceeding of the 55th PA Conference*, 2012b.
Reference Source

*Hitchman GM: **Testing the Psi mediated instrumental response theory using an implicit Psi task**. Doctoral Thesis, University of Northampton, England, 2012.
Reference Source

Honorton C, Ferrari DC: **"Future telling": A meta-analysis of forced-choice precognition experiments, 1935–1987.** *J Parapsychol.* 1989; **53**: 281–308.
Reference Source

Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, *et al.*: **Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index?** *Psychol Methods.* 2006; **11**(2): 193–206.
PubMed Abstract | Publisher Full Text

Jeffreys H: **Theory of probability**. OUP Oxford. 1998.
Reference Source

Jin ZC, Zhou XH, He J: **Statistical methods for dealing with publication bias in meta-analysis.** *Stat Med.* 2015; **34**(2): 343–360.
PubMed Abstract | Publisher Full Text

Judd CM, Gawronski B: **Editorial comment.** *J Pers Soc Psychol.* 2011; **100**(3): 406.
PubMed Abstract | Publisher Full Text

Judd CM, Westfall J, Kenny DA: **Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem.** *J Pers Soc Psychol.* 2012; **103**(1): 54–69.
PubMed Abstract | Publisher Full Text

Kahneman D: **Thinking, fast and slow**. New York: Farrar, Straus and Giroux, 2011.
Reference Source

Kaiser D: **How the hippies saved physics: science, counterculture, and the quantum revival**. New York: Norton, 2012.
Reference Source

Klauer KC, Musch J: **Affective priming: Findings and theories**. In J Musch & KC Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion*. Mahwah, NJ: Erlbaum, 2003; 7–49.
Reference Source

Lang PJ, Greenwald MK: **International affective picture system standardization procedure and results for affective judgments**. Gainesville, FL: University of Florida Center for Research in Psychophysiology, 1993.

LeBel EP, Peters KR: **Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice.** *Review of General Psychology.* 2011; **15**(4): 371–379.
Publisher Full Text

*Luke DP, Morin S: **Luck beliefs, PMIR, psi and the sheep-goat effect: A replication**. *Paper presented at the Society for Psychical Research 33rd International Conference*, University of Nottingham, 2009; 12–13.
Reference Source

*Luke DP, Delanoy D, Sherwood SJ: **Psi may look like luck: Perceived luckiness and beliefs about luck in relation to precognition.** *J Soc Psych Res.* 2008a; **72**: 193–207.
Reference Source

*Luke DP, Roe CA, Davison J: **Testing for forced-choice precognition using a hidden task: Two replications.** *J Parapsychol.* 2008b; **72**: 133–154.
Reference Source

*Macadam M: **Retrocausation and the sheep-goat effect: Challenging the idea that New Zealand is a nation of sheep**. Unpublished Honours Dissertation. University of Victoria, Wellington, New Zealand, 2011.

*Maier MA: **OrchOr Model of Consciousness: Experimental Evidence Part I**. *Paper presented at the TSC Arizona and personal communication*, 2012.

Maier MA, Buechner VL: **Time and consciousness**. In M Nadin (Ed.), *Anticipation Across Disciplines.* Berlin: SpringerVerlag, 2015.
Reference Source

*Maier MA, Büchner VL, Kuhbandner C, *et al.*: **Feeling the future again: Retroactive avoidance of negative stimuli.** *J Conscious Stud.* 2014; **21**(9–10): 121–152.
Reference Source

May EC, Marwaha SB, (Eds.): **Extrasensory Perception: Support, Skepticism, and Science**. Vol. 2: *Theories of Psi*. Santa Barbara, CA. Praeger, 2015.
Reference Source

*Milyavsky M: **Failure to replicate Bem (2011) Experiment 9**. Unpublished raw data, Hebrew University of Jerusalem, Jerusalem. Reported in Galak *et al.* 2012, 2010.

Morey RD, Rouder JN: **BayesFactor: Computation of Bayes factors for common designs**. R Package version 0.9.9. 2014.

*Morris B: **Precognitive habituation**. Daryl Bem's personal communication. 2004.

Mossbridge J, Tressoldi P, Utts J: **Predictive physiological anticipation preceding seemingly unpredictable stimuli: a meta-analysis.** *Front Psychol.* 2012; **3**: 390.
PubMed Abstract | Publisher Full Text | Free Full Text

Mossbridge JA, Tressoldi P, Utts J, *et al.*: **Predicting the unpredictable: critical analysis and practical implications of predictive anticipatory activity.** *Front Hum Neurosci.* 2014; **8**: 146.
PubMed Abstract | Publisher Full Text | Free Full Text

Mossbridge JA, Tressoldi P, Utts J, *et al.*: **We did see this coming: Response to "We should have seen this coming" By D Sam Schwarzkopf**. 2015.
Reference Source

*Moulton S, Kosslyn: **Retrocausal mere exposure**. *Presented at Psi and Psychology: The Recent Debate*. Harvard April 2011, 2011.

Orwin RG: **A fail-safe *N* for effect size in meta-analysis.** *J Educ Stat.* 1983; **8**(2): 157–159.
Publisher Full Text

*Parker A, Sjödén B: **Do some of us habituate to future emotional events?** *J Parapsychol.* 2010; **74**(1): 99–115.
Reference Source

*Pedersen JC, Shepardson SK, Lemka ZR, *et al.*: **Psi ability and belief: A replication of Bem (2011)**. *Poster presented at the 13th annual meeting of the Society of Personality and Social Psychology*, San Diego, CA, 2012.

*Platzer C: **Failure to replicate Bem (2011) Experiment 9**. Unpublished raw data, University of Mannheim, Mannheim, Germany, 2012.

*Popa IL, Batthyany A: **Retrocausal Habituation: A Study on Time-Reversal Effects in the Human Information Processing**. *Paper presented at the Cognitive Science Conference*, Bratislava, 2012.
Reference Source

*Rabeyron T: **Retro-priming, priming, and double testing: psi and replication in a test-retest design.** *Front Hum Neurosci.* 2014; **8**: 154.
PubMed Abstract | Publisher Full Text | Free Full Text

Radin DI: **Unconscious perception of future emotions: An experiment in presentiment.** *J Sci Explor.* 1997; **11**: 163–180.

Radin DI: **Entangled minds: Extrasensory experiences in a quantum reality**. New York, NY: Paraview Pocket Books, 2006.
Reference Source

*Ritchie SJ, Wiseman R, French CC: **Failing the future: three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect.** *PLoS One.* 2012; **7**(3): e33423.
PubMed Abstract | Publisher Full Text | Free Full Text

*Robinson E: **Not Feeling the Future: A Failed Replication of Retroactive Facilitation of Memory Recall.** *J Soc Psych Res.* 2011; **75**(904): 142–147.
Reference Source

*Roe C, Grierson S, Lomas A: **Feeling the future: two independent replication attempts**. *Parapsychological Association 55th Annual Convention,* Durham, North Carolina, 09–12 August 2012. Durham, North Carolina, USA: Parapsychological Association, 2012; 52–53.
Reference Source

Rosenthal R: **The "file drawer problem" and tolerance for null results.** *Psychol Bull.* 1979; **86**(3): 638–641.
Publisher Full Text

Rosenthal R, Rubin DB: **Interpersonal expectancy effects: the first 345 studies.** *Behav Brain Sci.* 1978; **1**(3): 377–415.
Publisher Full Text

Rothstein HR, Sutton AJ, Borenstein M: **Publication bias in meta analysis: prevention, assessment and adjustments**. West Sussex, England: Wiley, 2005.
Reference Source

Rouder JN, Morey RD: **A Bayes factor meta-analysis of Bem's ESP claim.** *Psychon Bull Rev.* 2011; **18**(4): 682–689.
PubMed Abstract | Publisher Full Text

Rouder JN, Speckman PL, Sun D, *et al.*: **Bayesian *t* tests for accepting and rejecting the null hypothesis.** *Psychon Bull Rev.* 2009; **16**(2): 225–237.
PubMed Abstract | Publisher Full Text

Rücker G, Schwarzer G, Carpenter JR, *et al.*: **Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis.** *Biostatistics.* 2011; **12**(1): 122–142.
PubMed Abstract | Publisher Full Text

*Savitsky K: Cited in Bem: **Precognitive Habituation: Replicable Evidence for a Process of Anomalous Cognition**. *Paper presented at the 46th Annual Convention of the Parapsychological Association*, Vancouver, BC August 2–4, 2003.
Reference Source

*Savva L, French CC: **Is there time-reversed interference in Stroop-based tasks**? *The Parapsychological Association 45th Annual Convention, Proceedings of the Presented Papers*, 2002; **66**(3): 194–205.
Reference Source

*Savva L, Child R, Smith MD: **The precognitive habituation effect: An adaptation using spider stimuli**. *Paper presented at the 47th Annual Convention of the Parapsychological Association*, Vienna, Austria, 2004.
Reference Source

*Savva L, Roe C, Smith MD: **Further testing of the precognitive habituation effect using spider stimuli**. *Paper presented at the Parapsychological Association, 48th August 11th – 14th* , 2005.

Schimmack U: **The ironic effect of significant results on the credibility of multiple-study articles.** *Psychol Methods.* 2012; **17**(4): 551–566.
PubMed Abstract | Publisher Full Text

Schwarzer G, Carpenter J, Rucker G: **Empirical evaluation suggests Copas selection model preferable to trim-and-fill method for selection bias in meta-analysis.** *J Clin Epidemiol.* 2010; **63**(3): 282–288.
PubMed Abstract | Publisher Full Text

Schwarzer G, Carpenter J, Rucker G: **Metasens. Advanced statistical methods to model and adjust for bias in meta-analysis R-package**. 2014.
Reference Source

Schwarzkopf DS: **We should have seen this coming.** *Front Hum Neurosci.* 2014; **8**: 332.
PubMed Abstract | Publisher Full Text | Free Full Text

Sheehan DP, (Ed.): **Frontiers of time: Retrocausation—experiment and theory**. AIP Conference Proceedings, San Diego, California. Melville, New York: American Institute of Physics, 2006; **1408**.
Reference Source

Sheehan DP, (Ed.): **Quantum Retrocausation—theory and experiment**. AIP Conference Proceedings, San Diego, California. Melville, New York: American Institute of Physics, 2011; **863**.
Reference Source

*Simmonds-Moore CA: **Exploring the Relationship between the synaesthesias and anomalous experiences**. Unpublished final report to the Bial foundation, 2013.
Reference Source

Simmons JP, Nelson LD, Simonsohn U: **False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant.** *Psychol Sci.* 2011; **22**(11): 1359–1366.
PubMed Abstract | Publisher Full Text

Simonsohn U, Nelson LD, Simmons JP: ***P*-Curve: a key to the file-drawer.** *J Exp Psychol Gen.* 2014a; **143**(2): 534–547.
PubMed Abstract | Publisher Full Text

Simonsohn U, Nelson LD, Simmons JP: ***p*-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results.** *Perspect Psychol Sci.* 2014b; **9**(6): 666–681.
PubMed Abstract | Publisher Full Text

Snodgrass S: **Examining retroactive facilitation of recall: an adapted replication of Bem (2011, Study 9) and Galak and Nelson (2010)**. 2011.
Publisher Full Text

Stanley TD: **Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection.** *Oxf Bull Econ Stat.* 2008; **70**(1): 103–127.
Publisher Full Text

Stanley TD, Doucouliagos H: **Meta-regression approximations to reduce publication selection bias.** *Res Synth Methods.* 2014; **5**(1): 60–78.
PubMed Abstract | Publisher Full Text

Stanley TD, Doucouliagos H: **Neither fixed nor random: weighted least squares meta-analysis.** *Stat Med.* 2015; **34**(13): 2116–2127.
PubMed Abstract | Publisher Full Text

*Starkie A: **Retroactive habituation: Exploring the time reversed amygdalar response to pictures of facial affect**. Dissertation presented at the Liverpool Hope University, 2009.
Reference Source

Sterne JAC, Egger M: **Regression methods to detect publication and other bias in meta-analysis**. In HR Rothstein, AJ Sutton, A. & M. Borenstein M. [Eds.], *Publication bias in meta analysis: prevention, assessment and adjustments*, West Sussex, England: Wiley, 2005.
Publisher Full Text

*Subbotsky E: **Sensing the Future: Reversed Causality or a Non-standard Observer Effect?** *The Open Psychology Journal.* 2013; **6**: 81–93.
Publisher Full Text

Thorndike EL: **Animal intelligence: An experimental study of the associative processes in animals.** *Psychol Monogr.* 1898; **2**(4): i–109.
Publisher Full Text

*Traxler MJ, Foss DJ, Podali R, *et al.*: **Feeling the past: the absence of experimental evidence for anomalous retroactive influences on text processing.** *Mem Cognit.* 2012; **40**(8): 1366–72.
PubMed Abstract | Publisher Full Text

*Tressoldi PE, Masserdotti F, Marana C: **Feeling the future: an exact replication of the Retroactive Facilitation of Recall II and Retroactive Priming experiments with Italian participants**. Università di Padova, Italy. Retrieved 05: 45, January 20, 2013, 2012.
Reference Source

*Tressoldi PE, Zanette S: **Feeling the future: an exact replication of the Retroactive Facilitation of Recall II and Precognitive Positive Detection experiments with Italian participants**. 2012.
Reference Source

Tressoldi P, Bem D, Rabeyron T, *et al.*: **Dataset 1 in: Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events.** *F1000Research.* 2015.
Data Source

van Assen MA, van Aert RC, Wicherts JM: **Meta-analysis using effect size distributions of only statistically significant studies.** *Psychol Methods.* 2015; **20**(3): 293–309.
PubMed Abstract | Publisher Full Text

*Vernon D: **Exploring the possibility of Precognitive Priming**. *Paper presented at the SPR annual conference*, University in Swansea, Wales, UK, 2013.

*Wagenmakers EJ, Wetzels R, Borsboom D, *et al.*: **An Agenda for Purely Confirmatory Research.** *Perspect Psychol Sci.* 2012; **7**(6): 632–638.
PubMed Abstract | Publisher Full Text

Wagenmakers EJ, Wetzels R, Borsboom D, *et al.*: **Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011).** *J Pers Soc Psychol.* 2011; **100**(3): 426–432.
PubMed Abstract | Publisher Full Text

Wagner MW, Monnet M: **Attitudes of college professors toward extra-sensory perception.** *Zetetic Scholar.* 1979; **5**: 7–17.
Reference Source

Walach H, Lucadou WV, Römer H: **Parapsychological phenomena as examples of generalized nonlocal correlations—A theoretical framework.** *J Sci Explor.* 2014; **28**(4): 605–631.
Reference Source

*Watt C, Nagtegaal M: **Luck in action? belief in good luck, Psi-mediated instrumental response, and games of chance.** *J Parapsychol.* 2000; **64**(1): 33–52.
Reference Source

Zajonc RB: **Attitudinal effects of mere exposure.** *J Pers Soc Psychol.* 1968; **9**(2, Pt 2): 1–27.
Publisher Full Text

*Zangari W: **Replication of the retro-habituation effect.** *Personal communication.* 2006.

Open Peer Review

Current Referee Status: ✓ ✓

Version 1

Referee Report 21 December 2015

doi:10.5256/f1000research.7733.r11510

✓ **Paul Grigoriev**
Department of Medical Physics and Informatics, Crimea State Medical University, Crimea, Ukraine

The research article “Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events” by Daryl Bem, Patrizio Tressoldi, Thomas Rabeyron, Michael Duggan raises the fundamental problem: if “nonlocal” anticipation really exists. In his original research Bem (2011) constructed several experimental designs that could indicate time-reversed cognitive and emotional effects similar to well-known and approved in psychology (approach and avoidance, priming, habituation, facilitation of recall).

Bem’s innovation to experimental approach in “psi” research is quite prominent because he started to use several methods based not on so-called force-choice. Thus, implicit drives and motivators started playing a role in results. For example, in experiments on “retroactive approach and avoidance” not the fact of guessing by itself is estimated, but the percentage of pictures with some similar content that may actualize similar emotions and drives in most people.

More generally, it is proved that intuition and anticipation are more effective and frequent in situations with lack of time and/or information for decision-making (Vasilieva, 2006). In special computer games with unpredicted results based on usage of random number generators we can see notable effect in guessing, especially in first 20 attempts (Li, 1992) while interest is high enough.

In present article, when Bem’s results excluded, the effect size became lesser (combined effect size for attempted replications by other investigators is 0.06, $z = 4.16$, $p = 1.1 \times 10^{-5}$). I do not think as a majority of sceptics, that Bem could select best results anyhow. For my opinion, it may be because Bem’s participants were the students – mostly young people with higher sex drive according to age and social status. In our recent investigation (Grigoriev, Vasilieva, 2015) we got similar results (to Bem’s) within young men (they guessed erotic photos more frequently), but results were opposite for men of higher age (they “guessed” erotic pictures significantly rarer than expected).

Really, Bem based his research, in particular, on thesis that participants should be attracted by erotic pictures and move away from unpleasant pictures. Thus, generally we follow our drives or motivations when select photos like we could see the “images of the future” or some contented hints from the future. If we even accept this fantastic assumption, we should remember than each individual is characterized by own structure of needs and psychological defenses, more over they depend from actual psychological and physiological state. So, relatively small effect size may be caused also by total averaging of participants. If we select participants to be similar in some respect, for example, men in prison or students under severe risk of expelling from university, we get more effect size (Grigoriev, Vasilieva, 2015), because real majority of participants were in similar and specific stress situations.

Underlying hypothesis about retroactive influence of future events on psyche and physiological state of individual seems to be reasonable in general, although we know nothing about physical mechanisms of such influence except of some metaphors from quantum physics like “nonlocality” on “entanglement”… Still entangled particles are interconnected not only in space, but also in time… Again, there is a place for strong discussion between different physicists. We do not know actually, if entanglement could be enough lasting and existing at high temperatures, and consist enough amount of information within particles like photons, for example.

Nevertheless, an argument that it is not worth to study something, only because nobody knows the mechanism of phenomenon, seems to be improper. At least, on the stage of exploratory researches. Moreover, skeptics’ arguments that anticipation of unpredictable events has no any sense for an organism is unqualified. Theory of functional systems of Pyotr Anokhin (Anokhin, 1973) is one of keystones of biology, like the meaning of anticipation (Lomov and Surkov, 1980; Lomov, 1984) for psychological adaptation.

We should care only about reproducibility. Yes, there are some difficult problems with reproducibility in investigations of psi in general. Although I do not think that retroactive influence of future events is “psi” at all. We do not know delicate physical and psychological mechanism that may work and provide quite small, but significant effect-size. One could offer at least several ways to overcome such weak reproducibility: 1. We should use longitude (during many days and repeats) for every participant and take into account different variables like physiological state and outer influences like weather, social events etc. 2. Parallel registration of processes in nervous system (McCraty, 2004; Mossbridge et al, 2012; Bierman & Scholte, 2002) to distinguish specific patterns in physiological processes when anomalous perception of unpredictable future events is successful. 3. Higher motivation on success or failure avoidance in participants. 4. Taking into account specific needs and state of every participant that may cause differences of unconscious setting to get the specific results.

Of course, this research is still quite odd for traditional psychology and may need some extraordinary evidences. At least authors provide us by enough statistical evidences of effects. The sources of possible artifacts were reviewed on the stages of data collecting, experimental methods and statistical analysis. It is worth to mention, that both Bayesian analysis and effect-size with statistical significance indices are enough high even after elimination of all “suspicious” data.

For my opinion, article may be published, because obtained positive results about possibility of anomalous retroactive influence reflect some unknown nonlocal mechanism that acts through time and can be felt by psyche and/or organism. It sounds weird; but we should remember every revolution in science had started from such odd facts, that incorporated then into the new theory…

And, as for the problem of weak reproducibility of results, the main strategy here should be the searching for the highly reproducible factors that cause variability of phenomenon. When we discover the reasons of nonreproducibility – thus we could understand a structure of phenomenon better.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 17 December 2015

doi:10.5256/f1000research.7733.r11511

✓ **Ina Vitalevna Vasileva**

Department of General and Social Psychology, Tyumen State University, Tyumen, Russian Federation

The article describes the results of meta-analysis 90 experiments performed to check is whether the anomalous anticipation of random future events exist. These results were obtained in 33 different laboratories located in 14 countries and involved 12.406 participants. Daryl Bem initially constructed and realized 4 types of experiments: 1. Precognitive Approach and Avoidance; 2. Retroactive Priming; 3. Retroactive Habituation; 4. Retroactive Facilitation of Recall. In each type of experiment, an influence of time-reserved effects was checked. An effect-size in researches included is enough to support the hypothesis about anomalous anticipation of random future events.

Online experiments revealed less successful comparing with traditional experiments with experimenter. Authors suppose that lesser effect-size in online-experiments is caused by inability to control all experimental conditions. I believe that it is an adequate explanation of differences in effect in online and laboratory experiments.
Authors discussed a possible influence of experimenter's belief in psi. It agrees with some hypothetical ideas of "quantum consciousness". Unfortunately, for three time-reserved experiments (1. Precognitive Approach and Avoidance; 2. Retroactive Priming; 3. Retroactive Habituation) such characteristics as valence and arousal for erotic and unpleasant photographs (The International Affective Picture System) were not described. We could estimate obtained results differential characteristics of participants (gender, age, social status). Particularly, in our recent research (Grigoriev, Vasilieva, 2015) we have obtained the results concerning relationship between such characteristics as a gender, age, satisfaction of basic needs and specific of affective visual stimuli anticipation (similar to "Precognitive Approach and Avoidance" type of experiment described in article and also using stimuli from IAPS) in cohorts of students (men and women), law-abiding and convicted of violent crimes men.

Authors also discussed the differences in effect-size between «fast-thinking» and «slow-thinking» strategies. They suggest that fast/slow variable effects on result of psi. From one side I agree with this experimental fact and conclusion, because «fast-thinking», as authors say, «prevent conscious cognitive strategies», although still suppose that detailed research is necessary to continue in respect to these variables, because participants' strategies may be caused not only with the speed of operation with stimuli, but also with content of stimuli.

This article may be accepted for indexing because of its proper methodological and methodical level. Enough variables that may affect psi were taken in consideration.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Discuss this Article

Version 1

Author Response 19 Jan 2016

Patrizio Tressoldi, Dipartimento di Psicologia Generale, Università di Padova, Italy

Authors' Responses to Comments on Version 1

Reply to Lakens:

We appreciate Lakens' posting of his review of the version of the article that we originally submitted to *Frontiers in Psychology*. *Frontiers* has the unusual policy of rejecting any article that fails to receive unanimous endorsement from all reviewers. Two previous reviewers had endorsed our article for publication and the Associate Editor was preparing to publish it when a general editor of *Frontiers* insisted that she seek yet another reviewer. That other reviewer turned out to be Lakens. Our article was then rejected and we never had the opportunity to revise our article in response to his review.

Therefore, in preparing our submission to *F1000Research*, we tried to be responsive to Lakens' concerns. In particular, we added to our analysis several additional statistical procedures that tested for selection bias in our database. As we note in our published article, only one of the 9 tests, the PET analysis Lakens cited in his review, included the zero effect size within its 95% confidence interval. We are pleased that Lakens now agrees that his original emphasis on that one analysis was too conservative. Even though we continue to disagree with his conclusions about our results, we appreciate both his respectful tone and the due diligence he displayed in his review and correspondence with us.

One of Lakens' criticisms of our article was the "lack of a theoretical framework." Ironically, psi researchers are simultaneously criticized for presenting psi data without an explanatory theory and for proposing theories of psi before establishing that they have data requiring explanation. Because our article is a meta-analysis, the emphasis is necessarily on the data. Note, however, that our General Discussion section lists several recent references that discuss substantive theoretical models that go well beyond metaphor and hand-waving in trying to accommodate psi phenomena.

Reply to Schwarzkopf:

We thank Schwarzkopf for spotting the error in the Abstract reporting the Bayes Factor (BF) for our database. The correct value (5.1×10^9) was the one displayed in our Table 1. We have now corrected that error in Version 2 of our article.

In Version 2 of our article, we have also adopted Schwarzkopf's suggestion that we report the robustness of the Bayes Factor across different ranges of the scaling factor *r* for the 69 independent replications of Bem's experiments. As noted there, the BF ranges from 3,853 when *r* is set to 0.1 (the value shown in Table 1) to a low of 992 when *r* is set to 1.0.

Regarding the outcome differences between Bem's original experiments and those reported in the meta-analysis, we discuss at length the differences between fast- and slow-thinking protocols reported in Table 2. Although the attempted replications of Bem's slow thinking experiment on retroactive recall do not achieve statistical significance, the *exact* replications of that protocol do. More generally, *p*-curve analysis estimates the true overall effect size of the independent replications of Bem's experiments to be 0.24, virtually identical to the effect size of his original experiments: 0.22.

Finally, as we note in both Versions 1 and 2 of our article, we agree with Schwarzkopf that research should now move to independent preregistered experiments, and current efforts in that direction are now underway.

Reply to Bierman:

Bierman's comment, with its bold headline "**Failed Replication**" is the most puzzling of the published comments. It, too, focuses on the difference in the effect sizes observed in the meta-analysis and those observed in Bem's original experiments.

We were already aware of Bierman's critique because he had previously posted it to a private online Listserv forum of psi researchers. When we queried him at that time about why he hadn't discussed the crucial *p*-curve analysis—which, as noted above, affirms the comparability of our meta-analytic results and Bem's original results—he acknowledged that he had not read about our *p*-curve analysis because he had not read beyond page 9 of our 22-page article before preparing his comments. He also acknowledged—and apologized for—overlooking the summary of that same analysis in our Abstract.

We understand that postings to Listserv forums are often quite informal and frequently contain errors of omission and commission. As with email messages generally, the "Send" button is often pressed prematurely. But what is puzzling and discomforting about Bierman's comment in this current forum is that he has simply reproduced his original posting without correcting his omission of our *p*-curve analysis.

P-curve analysis is quite new, and its validity, utility, implementation, and interpretation are certainly open to challenge. As we note in Version 2 of our article, the algorithm for performing the analysis is already in its fourth iteration. Bierman might well find grounds to question our use of it. But we respectfully ask that he read about it before stating "Failed Replication."

Competing Interests: These are comments submitted on behalf of all authors

Reader Comment 08 Jan 2016

D. Sam Schwarzkopf, UCL Experimental Psychology & Institute of Cognitive Neuroscience, UK

Another, brief comment after I had some additional time to look at the study:

The abstract states that the Bayes factor in support of the alternative hypothesis for the full meta-analysis is 1.4 x 10^9. This is actually lower than what is reported in the text of the paper where it is given at 5.1 x 10^9, which is the Bayes factor obtained with a Cauchy scaling factor of r=0.1. For the smaller scaling factor tested, r=1, the Bayes factor is 2 x 10^9, so still larger than what is reported in the abstract. Presumably this is a typo that should be corrected?

More importantly, how does the robustness analysis look when Bem's own data are excluded? For the standard prior chosen for the main analysis the BF is 3,853, so approximately one millionth of what it is when his data are included. Considering that there is such a stark discrepancy in the effect sizes of his experiments (mean g=0.22) and all the other experiments (mean g=0.08; something Dick Bierman's comment above also discusses), a difference that is in fact significant ((t(88)=2.75, p=0.007, BF10=6.045), and the general controversy around these experiments, more attention should be paid to independent (ideally, preregistered) experiments.

Competing Interests: No competing interests were disclosed.

Reader Comment 06 Jan 2016

Daniel Lakens, Eindhoven University of Technology, Netherlands

Review of Feeling the Future: A Meta-analysis of 90 Experiments on the Anomalous Anticipation of Random Future Events by Bem, Tressoldi, Rabeyron, & John Duggan.

This is a review I wrote for Frontiers, where the manuscript was not accepted for publication. I would now change the stress I put on the PET analysis - it may be too conservative. But the other points remain, and should be sufficient to not take the current meta-analysis as scientific evidence of pre-cognition.

I have previously looked at this meta-analysis for a blog post I've written (Lakens, 2014, http://daniellakens.blogspot.nl/2014/05/a-pre-publication-peer-review-of-meta.html). I had a very professional exchange with the authors, which was pleasurable and interesting, and led the authors to correct the mistakes I pointed out and answer some questions I had. I thought it was interesting to peer review an article that had been posted in a public depository.
Now, my task is different, according to the Frontiers review guidelines: "The mandate for review editors is to ensure that the results are valid, the analysis is flawless and the quality as high as possible" In the review below, I take this task very seriously, and regret to have to conclude that the results are not valid, the analyses are flawed in many respects, and the quality is too low for this meta-analysis to be part of the scientific literature. I believe this manuscript should be rejected.
Let me be clear that I would have no problem with a well-performed meta-analysis of this literature, regardless of whether it would show a meta-analytic effect size estimate that differed from zero or not. In science, we distinguish between statistical inferences and theoretical inferences (e.g., Meehl, 1990). Even if a meta-analysis would lead to the statistical inference that there is a signal in the noise, there is as of yet no compelling reason to draw the theoretical inference that psi exists, due to the lack of a theoretical framework as acknowledged by the authors. So, a meta-analytical effect size estimate that differs from zero would have to lead to a careful examination of possible confounds in the paradigms that have been used in this literature, and the studies that have been included in this meta-analysis. Such a careful examination has not been done. Therefore, the validity of the measures used to examine psi effects has not been established. Only 18 statistically significant effects have been observed in the last 14 years, as the literature search by the authors reveals, coming from only 7 labs. Only if confounds are sufficiently excluded (preferably in direct replications in different labs) can we start thinking about alternative explanations for the observed data, such as psi. In other words, even if there was robust evidence for a meta-analytic effect size estimate that differed from zero in our statistical inferences, we are far removed from being able to draw any theoretical inferences. If research on psi has demonstrated anything, it is that when you lack a theoretical model, scientific insights are gained at a painstakingly slow pace, if they are gained at all.
Before we accept the conclusion in the manuscript (and abstract) that there is an overall effect size of 0.09, we need to check whether the meta-analysis has been performed adequately, and whether bias has influenced this meta-analytic effect size estimate.

Dealing with publication bias.

The authors use Begg and Mazumdar's rank correlation test to examine publication bias, stating that: "The preferred method for calculating this is the Begg and Mazumdar's rank correlation test, which calculates the rank correlation (Kendall's tau) between the variances or standard errors of the studies and their standardized effect sizes (Rothstein, Sutton & Borenstein, 2005)."
I could not find this recommendation in Rothstein *et al.*, 2005). From the same book, chapter 11, p. 196,

about the rank correlation test:
Sterne and Egger (Chapter 6) caution against using the test unless the meta-analysis includes a range of study sizes, including at least one of 'medium' size. Otherwise, the result will be driven primarily by noise. They also note that the test has low power unless there is severe bias, and so a non-significant tau should not be taken as proof that bias is absent (see also Sterne *et al*., 2000, 2001b, c)
Similarly, from the Cochrane handbook of meta-analyses (http://handbook.cochrane.org/chapter_10/10_4_3_1_recommendations_on_testing_for_funnel_plot_asymme):
"The test proposed by Begg and Mazumdar (Begg 1994) has the same statistical problems but lower power than the test of Egger *et al*., and is therefore not recommended."
When the observed effect size is tiny (as in the case of the current meta-analysis), just a small amount of bias can yield a small meta-analytic effect size estimate that is statistically different from 0. In other words, whereas a significant test result is reason to worry, a non-significant test result is not reason not to worry.
Also note how the Cochrane handbook suggests Egger's test to be a superior alternative (I will show below that meta-regressions shows that correcting for publication bias, the meta-analytic effect size is not significantly different from 0).
The authors also report the trim-and-fill method to correct for publication bias. It is known that when publication bias is induced by a *p*-value boundary, rather than an effect size boundary, and there is considerable heterogeneity in the effects included in the meta-analysis, the trim-and-fill method might not perform well enough to yield a corrected meta-analytic effect size estimate that is close to the true effect size (Peters, Sutton, Jones, Abrams, & Rushton, 2007; Terrin, Schmid, Lau, & Olkin, 2003).

Similarly, from the Cochrane handbook (http://handbook.cochrane.org/chapter_10/10_4_4_2_trim_and_fill.htm):

"The trim and fill method requires no assumptions about the mechanism leading to publication bias, provides an estimate of the number of missing studies, and also provides an estimated intervention effect 'adjusted' for the publication bias (based on the filled studies). However, it is built on the strong assumption that there should be a symmetric funnel plot, and there is no guarantee that the adjusted intervention effect matches what would have been observed in the absence of publication bias, since we cannot know the true mechanism for publication bias. Equally importantly, the trim and fill method does not take into account reasons for funnel plot asymmetry other than publication bias. Therefore, 'corrected' intervention effect estimates from this method should be interpreted with great caution. The method is known to perform poorly in the presence of substantial between-study heterogeneity(Terrin 2003, Peters 2007). Additionally, estimation and inferences are based on a dataset containing imputed intervention effect estimates. Such estimates, it can be argued, inappropriately contribute information that reduces the uncertainty in the summary intervention effect."

The authors nevertheless assume the trim-and-fill methods provides an indication that publication bias is not a problem, and present a 'corrected' effect size estimate. However, the corrected effect size is completely untrustworthy.

Better tests for publication bias
The authors present fail-safe N, Begg and Mazumdar's rank correlation test, and the trim-and-fill method. All these tests should be removed from the manuscript, or at the very best added to supplementary materials and only cursory discussed within the manuscript itself, pointing out they do not tell us much. I know they are often used, but we can hardly use the continued mistakes of others as an excuse to follow scientifically invalid practices. The author's justification that trim-and-fill 'is currently the most common method for estimating the number of missing studies in a meta-analysis' is therefore akin to admitting you

are using a statistical tool that is known not to work well, because most others are using that technique as well. Researchers often attempt to justify the inclusion of these measures when they give favorable results, but no justification is possible. I think reporting the fail-safe *N* in the abstract is very misleading, especially given that most people in the world (psychologists included) don't understand statistics very well, and will misinterpret it. After removing rank correlation, trim-and-fill, and fail-safe N, present contour enhanced funnel plots, instead of the normal funnel plots in the manuscript (e.g., Figure 1), also in the appendices.
While the authors are interested in using more novel tests (e.g., *p*-curve analyses) they do not use updated versions of Eggers regression to examine publication bias, namely PET-PEESE meta-regression, even though this seems to be the best test to examine publication bias we currently have. This approach is based on first using the precision-effect test (PET, Stanley, 2008) to examine whether there is a true effect beyond publication bias, and then follow up on this test (if the confidence intervals for the estimate exclude 0) by a PEESE (precision-effect estimate with standard error, (Stanley and Doucouliagos, 2007) to estimate the true effect size.
In the R code where I have reproduced the meta-analysis, I have included the PET-PEESE meta-regression. The results are clear: the estimated effect size without publication bias is 0.008, and the confidence intervals around this effect size estimate do not exclude 0. In other words, there is no good reason to assume that anything more than publication bias is going on in this meta-analysis.

```
Call:
lm(formula = d.all ~ d.se.all, weights = 1/d.v.all)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-2.0032 -0.7251  0.2041  0.6933  1.6760

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.008344   0.021571   0.387  0.69984
d.se.all    0.548656   0.178451   3.075  0.00281 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8882 on 88 degrees of freedom
Multiple R-squared:  0.097,Adjusted R-squared:  0.08674
F-statistic: 9.453 on 1 and 88 DF,  p-value: 0.002808

> confint(SE.all)
                  2.5 %     97.5 %
(Intercept) -0.03452504 0.05121256
d.se.all     0.19402280 0.90328951
```

Using the best technique available today to examine the correlation between 'study sizes and effect sizes' (the authors should use 'standard errors and effect sizes') we thus have to conclude the positive overall effect size is due to publication bias. I think this article should clearly specify that there is no effect of psi in a meta-analysis over 90 studies in the abstract, and that there is therefore no scientific evidence for the presence of psi. Publishing this article with any other conclusion in the abstract would be misleading. I understand the authors might need some time to adjust to this conclusion, now that they have been provided with a statistical technique they might not have heard of, but when they examine this technique in more detail, they will undoubtedly come to the conclusion that their meta-analysis indicates the most likely

meta-analytic effect size estimate their data provides is a psi effect of 0.
Perhaps it will help to realize that if psi had an effect size of Cohen's dz = 0.09, to have 90% power to examine an effect with an effect size estimate of 0.09, an alpha level of 0.05, and performing a two-sided t-test, you'd need 1300 participants. Only 1 experiment has been performed with a sufficiently large sample size (Galak, exp 7), and this experiment did not show an effect. Meier (study 3) has 1222 participants, and finds an effect at a significance level of 0.05. However, using a significance level of 0.05 is rather silly when sample sizes are so large (see http://daniellakens.blogspot.nl/2014/05/the-probability-of-p-values-as-function.html) and when we calculate a Bayes Factor using the *t*-value and the sample size, we see this results in a JZS Bayes Factor of 1.90 – nothing that should convince us.

```
library(BayesFactor)
1/exp(ttest.tstat(t=2.37, n1=1222, rscale = 0.707)[['bf']])

[1] 1.895727
```

I hope the authors will realize that a meta-analysis of primarily random noise mixed with publication bias can hardly be expected to provide convincing support of psi effects. The authors would do better to perform two or three pre-registered studies using a sample size of around 1300 participants, and repeatedly show there is an effect in the paradigm they use (even though this would not immediately mean the effect is related to psi).

Estimating the evidential value with *p*-curve and *p*-uniform.
The authors report two analyses to examine the effect size based on the distribution of *p*-values. These techniques are new, and although it is great the authors embrace these techniques, they should be used with caution. For example, a new version of the *p*-curve app has just been released, which uses a slightly different calculation of the critical tests, and slightly different labels for the tests that are performed. The R code for the *p*-uniform technique has only recently been made available to a select number of scholars (I have access to the code) but lacks documentation and has not been carefully compared with other techniques, such as *p*-curve, PET-PEESE, etc. – just to show these techniques are still being developed. The new test of the *p*-curve app return a statistically significant effect when testing for right skew, or evidential value. However, it now also includes an exploration of how much this test result depends on a single *p*-value, but plotting the significance levels of the test if the *k* most extreme *p*-values are removed. As we see in the graph below, the test for evidential value returns a *p*-value above 0.05 after excluding only 1 *p*-value, which means we cannot put a lot of confidence in these results.

I also think it is important to note that I have already uncovered many coding errors in a previous blog post (Lakens, 2014), even though the authors note that 2 authors independently coded the effect sizes. I feel I could keep pointing out more and more errors in the meta-analysis (instead, I will just repeatedly recommend to include a real statistician as a co-author), but let's add one to illustrate how easily the conclusion in the current *p*-curve analysis changes. The authors include Bierman and Bijl (2013) in their spreadsheet. The raw data of this experiment is shared by Bierman and Bijl (and available at: https://www.dropbox.com/s/j44lvj0c561o5in/Main%20datafile.sav - another excellent example of open science), and I can see that although Bierman and Bijl exclude one participant for missing data, the reaction times that are included in the meta-analysis are not missing. Instead, the data from the Human Information Processing questionnaire participants performed after the reaction time study is missing. Indeed, in the master thesis itself (Bijl & Bierman, 2013 http://www.uniamsterdam.nl/D.J.Bierman/PUBS/2013/Bijl_Bierman_PA2013.docx), all reaction time data is included. If I reanalyze the data, I find the same result as in the master thesis:

I don’t think there can be much debate about whether all reaction time data should have been included, and I think that the choice to report reaction time data from 67 instead of 68 participants in one of those tiny sources of bias that creep into the decisions researchers almost unconsciously make (after all, the results were statistically significant from zero regardless of the final choice). However, for the *p*-curve analysis (which assumes authors stop their analysis when *p*-values are smaller than 0.05) this small difference matters. If we include t(67)=2.11 in the *p*-curve analysis instead of *t*(67)=2.59, the new *p*-curve test no longer indicates the studies have evidential value.

On a final note, it should be pointed out that the correct degrees of freedom in the results used by Bem and colleagues should be 66 and not 67 (because they use the data of 67, not 68 participants). Bierman and Bijl (2013) correctly report the degrees of freedom themselves, but Bem and colleagues initially had included the correct test (e.g., in the version of the meta-analysis I reviewed on my blog). They later followed Bierman and Bijl (2013) by using the test performed on 67 instead of 68 participants, changing the *t*-value and *p*-value, but forgetting to change the degrees of freedom. I want to point out that such a strategic selection of the lowest possible *p*-value (by whoever is finally responsible for introducing this bias) is not in line with an assumption of the *p*-curve test (which assumes people stop selectively reporting whatever gives them a p < 0.05). This example demonstrates this is not always the case.

One problem with new techniques is that researchers might not understand them. The authors have asked the authors of the *p*-uniform papers to perform an analysis on their data, and conclude: “For our database, *p*-uniform confirms that there is, in fact, a significant effect in our database (p = .005) and that there is no evidence for selection bias (p = .857).” I delved in this new technique (and would like to thank Robbie van Aert and Marcel van Assen for their help), and quickly learned that in principle the test against a uniform distribution should be the same for a *p*-curve and *p*-uniform analysis. The reason they differ in the manuscript is because the two tests were calculated based on different assumptions. *P*-uniform is calculated based on the idea that all reported tests were one-sided. The *p*-values for all tests are therefore half as large in the *p*-uniform test (e.g., p = 0.007) than in the *p*-curve test (e.g., p = 0.014). The test is performed on a larger subset of the data (29 instead of 17 studies). The differences between the tests are not based on the differences in the data, but in the way the data is used and the decisions about the test that are made. I think this nicely illustrates the authors often lack sufficient understanding of the statistical techniques they use. They need to cooperate with a real statistician if the want to perform a meta-analysis that should be taken seriously by the researcher community.

The new version of the *p*-curve test shows there is evidential value. But instead of mindlessly interpreting the *p*-values we get from the analyses, let’s first look at the plot of our data. We see a very weird *p*-value distribution that would not be predicted. There are many more *p*-values between 0.01-0.02 then between 0.00-0.01. Under typical circumstances, we should see many, many more *p*-values below 0.01 than between 0.01-0.02 (e.g., Lakens, 2014).

Remember that *p*-curve is a relatively new technique. For many tests we use (e.g., the *t*-test) we first perform assumption checks. In the case of the t-test, we check the normality assumption. If data isn’t normally distributed, we cannot trust the conclusions from a *t*-test. I would severely doubt whether we can trust the conclusion from this *p*-curve. Regardless of whether the *p*-curve tells us there is evidential value or not, the *p*-curve doesn’t look like a ‘normal *p*-value distribution’. Consider the *p*-curve analysis as an overall F-test for an interaction. The *p*-curve tells us there is an effect, but if we then perform the simple effects (looking at *p*-values between 0.00-0.01, and between 0.01-0.02) our predictions about what these effects look like is not confirmed. Again, this is just my own interpretation of the *p*-curve test, and it will be useful to see how this test develops. For now, I just want to conclude it is debatable whether the conclusion there is an effect has passed the *p*-curve test for evidential value (I would say it has not), and passing the test is not immediately a guarantee there is evidential value.

Estimating the true effect size with *p*-curve and *p*-uniform.

Robbie van Aert, Marcel van Assen, and Jelte Wicherts were kind enough to share a recent commentary article that is under review ("Three reservations on *p*-curve for meta-analysis: A comment on Simonsohn, Nelson, and Simmons (2014)". In this article, they examine the performance of *p*-uniform and *p*-curve when there is variability in the effect size estimates. Their abstract:

"Because evidence of publication bias in psychology is overwhelming, it is important to develop techniques that correct meta-analytic estimates for publication bias. Simonsohn, Nelson, and Simmons (2014) show that *p*-curve and the methodology on which it is based have great promise for providing accurate meta-analytic estimates in the presence of publication bias. However, we show that *p*-curve behaves erratically in some situations. Moreover, we show and explain that, as opposed to statements in Simonsohn *et al*. (2014), *p*-curve overestimates effect size under moderate to large heterogeneity, and may yield unpredictable bias when researchers employ *p*-hacking . We therefore conclude that estimates of *p*-curve and *p*-uniform (van Assen, van Aert, and Wicherts, 2014), a method based on the same principles but differing from *p*-curve in implementation, should be interpreted cautiously in case of evidence of heterogeneous effect sizes or extensive *p*-hacking. Finally, we suggest ways to improve the accuracy of *p*-curve and *p*-uniform"

We need to know how much heterogeneity there is in the meta-analysis. Regrettably, the authors make no attempt to share or explain the heterogeneity in the manuscript. Please provide an index of the heterogeneity (e.g., *I2*) in every performed meta-analysis. Furthermore the author should attempt to explain this heterogeneity – they perform subsample analyses in their manuscript, but it is uncertain whether or which subsamples differ. Is the difference between slow-thinking and fast-thinking paradigms explaining the heterogeneity?
If my meta-analysis is correct, heterogeneity in the overall analysis is substantial:
Quantifying heterogeneity:
tau^2 = 0.0056; H = 1.31 [1.16; 1.49]; I^2 = 42.1% [25.4%; 55.1%]

```
Test of heterogeneity:
     Q d.f.  p-value
 153.77   89 < 0.0001
```

This has consequences for the use of *p*-curve and *p*-uniform to estimate effect sizes.
The authors state that: 'Simonsohn *et al*. (2014b) state that the accuracy of their effect-size estimate "does not rely on homogeneity of sample size or effect size. In all cases, *p*-curve is accurate and the other methods are not (p. 670)." This will without a doubt not be the first time a researcher has made a slightly boisterous claim about his own recently developed statistical technique that turns out to be false, and I don't think the authors need to rub this in by providing a direct quote. I'm confident future studies will point our heterogeneity is a problem. I observed this myself on my blog (http://daniellakens.blogspot.nl/2014/09/publication-bias-in-psychology-putting.html) and Aert *et al* (under review) provide more evidence for this. From Aert *et al*:
"The other columns, however, show that both *p*-curve and *p*-uniform overestimate the mean population effect size of .397 for moderate to large heterogeneity, and that this bias increases with larger heterogeneity."
and
"However, presently we do not yet recommend estimating effect size with *p*-curve or *p*-uniform when there is evidence of moderate to large heterogeneity, or when there are methodological, psychometric, or substantive reasons to expect effects to be relatively heterogeneous."

Although the manuscript by Aert *et al* is still under review, the take home message should be that both types of tests can overestimate the true effect size. The statement by the authors that ‘This implies that the higher estimate of .20 for the true effect size in our database is the correct one’ is completely unwarranted and should be removed. The authors should instead simply rely on the statistical technique that, according to our current understanding, is the most reliable, which is the PET estimate, as far as I am aware. *P*-curve and *p*-uniform are promising techniques, but it is too early to use these techniques to argue for a true effect in this meta-analysis.

Remember that almost all tests for publication bias work as a sensitivity test under many circumstances. That is, when they indicate problems (such as indicated by the *p*-curve analysis and the PET-analysis) we can be pretty sure there are real problems, but if they show there is an effect, we can hardly be certain a true effect exists. This is regrettable, but an inevitable consequence of publication bias.

The presence of bias

In the literature, a lot has been said about the fact that the low-powered studies reported in Bem (2011) strongly suggest there are an additional number of unreported experiments, or that the effect size estimates were artificially inflated by *p*-hacking (see Francis, 2012). The authors mention the following when discussing the possibility that there is a file-drawer (page 9):

“In his own discussion of potential file-drawer issues, Bem (2011) reported that they arose most acutely in his two earliest experiments (on retroactive habituation) because they required extensive preexperiment pilot testing to select and match pairs of photographs and to adjust the number and timing of the repeated subliminal stimulus exposures. Once these were determined, however, the protocol was “frozen” and the formal experiments begun. Results from the first experiment were used to rematch several of the photographs used for its subsequent replication. In turn, these two initial experiments provided data relevant for setting the experimental procedures and parameters used in all the subsequent experiments. As Bem explicitly stated in his article, he omitted one exploratory experiment conducted after he had completed the original habituation experiment and its successful replication.”

This is not sufficient. The power for his studies is too low to have observed the number of low *p*-values reported in Bem (2011) without having a much more substantial file-drawer, or *p*-hacking. It simply is not possible, and I will not accept vague statements about what has been reported. Where I would normally give researchers the benefit of the doubt (our science is built on this, to a certain extent) I cannot do this when there is a clear statistical indication that something is wrong. I also want to point out that, in addition to a high frequency of low *p*-values, the effect size estimates provided by Dr Bem are twice as large as the overall meta-analytic effect size estimate (see meta-analysis performed only on Dr Bem’s data below):

The funnel plot looks like this:

Both the average effect size by Dr Bem, the number of significant results despite low power, and the funnel plot lead to a single conclusion: I cannot trust the effect sizes in this set of studies. The signs of bias are clear. Note that my conclusions about the presence of bias are further strengthened by the PET analysis I have reported above which indicated there is no real effect of Psi after controlling for publication bias. The lack of an effect in the PET analysis remains true even after excluding all data contributed by Dr. Bem (in other words, publication bias is common and not unique to the work by Dr. Bem). However, Dr. Bem could have easily included remaining studies in the meta-analysis, or acknowledged that the performed analyses in many studies consisted of exploratory analyses with inflated Type 1 error rates. That the authors don’t question the results reported in Bem (2011) is a clear sign of bias. A similar (but less severe) sign of bias is the fact that a recent study by Rabeyron (2014, accepted March 2014) with a negative effect size, and revealing a failed replication of Rabeyron & Watts (2010) is not included in this meta-analysis or discussed in the manuscript (which was submitted in November 2014). I could go on (there are many stylistic sources of bias, the abstract being the best example), but I simply think there is more than enough reason to worry.

I would ask the authors to fill out the PRISMA statement (http://www.prisma-statement.org/statement.htm) instead of simply stating they complied to APA reporting standards, and add the PRISMA statement to the supplementary materials. I'm especially interested in how the authors have dealt with bias – it is clear that the authors are involved in this research, and have a lot to benefit from a positive outcome of this meta-analysis, which is known to bias the effect size estimate in meta-analyses.

Furthermore, although the authors note two individuals have independently coded the results, it was rather easy for me to identify many coding errors when calculating effect sizes (Lakens, 2014). I looked at only a subset of the studies, and not even very critically at those. That I nevertheless observed many errors leads me to the recommendation to contact a statistician, and ask this statistician to check all analyses.

I think the manuscript would have been better (but also have lead to completely opposite conclusions) when the authors would have asked someone with sufficient expertise to perform a meta-analysis to join the research team. Given that this manuscript has been going through the review process for a long time, and errors are still being discovered, I think additional checks on the data extraction and the performed analyses should be performed. Please note that I am not a statistician myself (I would say I know almost nothing about statistics) such that even my comments here should ideally be confirmed by a true expert. In case the peer review process for this manuscript would be continued, at the very minimum, it should be required that a real statistician with expertise in meta-analyses checks all analyses and assists the researchers in analyzing and reporting the data. I'd also appreciate it if the authors could provide all articles and raw data necessary to check every calculated effect size included in the meta-analysis in an online repository.

I understand the goal of Frontiers is to help authors improve their manuscript (I am an editor at Frontiers myself). However, in this instance, I believe the quality of the work is too low for this to be a fruitful approach, and I lack the confidence that the current authors will be able to produce a meta-analysis that is of sufficient quality and adequately addresses the bias that is so clearly present in this manuscript.

I feel that publishing this would hurt the scientific study of psi. For this research area to be taken seriously be scientists, it should make every attempt to be free from bias. I know many researchers in this field, among others Dr Tressoldi, are making every attempt to meet the highest possible standards, for example by publishing pre-registered studies (e.g., https://koestlerunit.wordpress.com/study-registry/registered-studies/). I think this is the true way forward (but I also think it is telling us something that if replications are performed, even by the original authors (e.g., Rabyeron, 2014) these studies consistently fail to replicate the original results). Publishing a biased meta-analysis stating in the abstract there is '"decisive evidence" in support of the experimental hypothesis' while upon closer scrutiny, the meta-analysis fails to provide any conclusive evidence of the presence of an effect (let alone support for the hypothesis that psi exists) would be a step back, rather than a step forward.

Conclusion

No researcher should be convinced by this meta-analysis that psi effects exist. I think it is comforting that PET meta-regression indicates the effect is not reliably different from 0 after controlling for publication bias, and that *p*-curve analyses do not indicate the studies have evidential value. However, even when statistical techniques would all conclude there is no bias, we should not be fooled into thinking there is no bias. There most likely will be bias, but statistical techniques are simply limited in the bias they can reliably indicate.

My biggest concern with respect to the current meta-analysis is not the small errors in the calculations of effect sizes or the *p*-curve, nor the use of many techniques that are widely believed to be outdated and inaccurate while interpreting these results in favor of the hypothesis, nor the lack of knowledge about some of the newer statistical techniques the authors use, but primarily the clear bias in the meta-analysis. Performing meta-analyses on biased data will not lead to reliable conclusions, and I have severe doubts this bias can be overcome. Psi effects are an important research area in the eyes of the general public. Let's not allow low quality work on psi to discredit the status of psi research in particular, and science in

general. Instead, we need better evidence before we attempt to draw meta-analytical conclusions about whether specific paradigms yield reliable effects.
If this review process continues (I don't believe it should, as detailed above), I think that based on my review, the abstract of the manuscript in a future revision should read as follows:
In 2011, the Journal of Personality and Social Psychology published a report of nine experiments purporting to demonstrate that an individual's cognitive and affective responses can be influenced by randomly selected stimulus events that do not occur until after his or her responses have already been made and recorded, a generalized variant of the phenomenon traditionally denoted by the term precognition (Bem, 2011). To encourage replications, all materials needed to conduct them were made available on request. We here report a meta-analysis of 90 experiments from 33 laboratories in 14 countries which yielded an overall effect size (Hedges' g) of 0.09, which after controlling for publication bias using a PET-meta-regression is reduced to 0.008, which is not reliably different from 0, 95% CI [-0.03; 0.05]. These results suggest positive findings in the literature are an indication of the ubiquitous presence of publication bias, but cannot be interpreted as support for psi-phenomena. In line with these conclusions, a p-curve analysis on the 18 significant studies did not provide evidential value for a true effect. We discuss the controversial status of precognition and other anomalous effects collectively known as psi, and stress that even if future statistical inferences from meta-analyses would result in an effect size estimate that is statistically different from zero, the results would not allow for any theoretical inferences about the existence of psi as long as there are no theoretical explanations for psi-phenomena.

Competing Interests: no competing interests

Reader Comment 09 Nov 2015

Dick Bierman, University of Amsterdam, University of Groningen, Netherlands

FAILED REPLICATION
In this article 'replication' is the most prominent word. The authors write:..... We agree. Rather than continuing to debate Bem's original experiments, we seek in our meta-analysis to answer the one question that most decisively trumps such disputes: Can independent investigators replicate the original experiments?........
The answer is obviously NO a because a comparison between the 'original' Bem studies and the replication efforts show that the results the former do significantly deviate from the results in the latter attempted replication. Although the simple analysis that I performed is not weighted for the different paradigms win the original study the p of 0.0037 is small enough to survive more subtle evaluations. Also there are only 2 studies in the whole database that have enough power to expect a significant outcome given the effect size that Bem now claims. It has been argued that one should never use underpowered studies in a meta-analysis. (Muncer S, Taylor S, Craigie M. Power dressing and meta-analysis: incorporating power analysis into meta-analysis. Journal of Advanced Nursing 2002;38(3):274-280.). What is needed now are *well-powere*d studies to confirm the claimed es of 0.06. I.e the sample sizes should be close to 750.
Interestingly simulations of questionable research practices for the GF-telepathy paradigm meta analysis show that the best explanation of the database is obtained if the originally claimed effect size of ~0.17 is reduced to ~0.06-0.07, the same value as now is claimed on the basis of these 'retrocausal' studies.(Bierman and Spottiswoode, in press)

Competing Interests: No competing interests were disclosed.

Reader Comment 02 Nov 2015

D. Sam Schwarzkopf, UCL Experimental Psychology & Institute of Cognitive Neuroscience, UK

I have read several versions of this manuscript before. A review by Daniel Lakens (published on his blog: http://daniellakens.blogspot.co.uk/2015/04/why-meta-analysis-of-90-precognition.html) identified indications that publication bias may skew these results. I leave it up to the reader to judge in how far his comments have been addressed in this new version.

I want to comment on a separate issue: the discussion of how Bayesian evidence is used to update one's belief about a finding (see paragraph starting with 'An opportunity to calculate...' in General Discussion). The authors rightly argue that the Bayes Factor is to be interpreted as evidence that the experimenter or reader can use to update their guestimate whether or not precognition is real. They say that skeptics of "Psi" (in this case specifically a study by Wagenmakers et al., 2011) have a very strong prior (10^20) against the existence of Psi and that even their "decisive" Bayes Factor is insufficient to shift that belief to convince them otherwise. Even after all this the skeptics' posterior odds are tilted far away from accepting the existence of Psi effects. (At this point I would like to point out a great blog post by Alexander Etz illustrating the distinction between statistical evidence and drawing a conclusion: http://alexanderetz.com/2015/11/01/evidence-vs-conclusions/)

The authors say "Clearly psi-proponents have their work cut out for them." What they do not say is what their prior belief is. Clearly, as self-identified "psi-proponents", they must set their prior odds to be somewhere in favour of Psi. Everybody is entitled to their own priors but some priors are probably more defensible than others. As I have repeatedly argued, not only in the context of Psi research but also in general terms, researchers should take into account the scientific plausibility of the effects they observe. They should predict what kinds of effect sizes are likely a priori and whether the observed effects are consistent with these predictions. I emphathise with the notion that this is hard in Psi research because the theoretical basis of these effects is vague at best. However, I also think it is particularly pertinent for claims of this magnitude.

And it is not an insurmountable problem. If the researchers believe in some quantum mechanistic effect or whatever kind of time-symmetric phenomenon they should come up with a prediction of what kinds of statistical effect sizes this is likely to produce in experiments of this kind (for instance, 12-18 trials with probability 0.5 per condition per subject in Bem's experiment 1). My hunch is that the expected effect size is several orders of magnitude smaller than the already miniscule effect size estimated by this meta-analysis.

Another approach could be to take the estimated effect size from this meta-analysis and try to simulate how this would manifest in the real world if the effect were true. If 100 people will make correct guesses for emotional events 51-53% of the time, how will this impact everyday events, such as profits by casinos? You'd think if a precognition effect were so easily measurable as these experiments imply we should see a considerable impact of that on the economy and the general survival of our species. Psi research is often motivated by anecdotal evidence of weird "anomalous" phenomena, like precognitive dreams or knowing the phone will ring with an important call just before it happens. How frequent are those events really and how consistent is that rate with cognitive biases compared to actual precognition?

The answers to these questions could actually inform educated prior odds that can tell us how likely it is that effects such as the ones observed in these experiments are really due to precognition. I don't know what the answer is but I think we can be confident that the odds against it are stronger than 1 to 1. In the absence of more informed prior odds, I think 10^20 is a pretty good bet.

Competing Interests: I am extremely skeptical of the claim that precognition exists - but if we announced our prior skepticism as competing interests we should declare conflicts of interest for every study we author or review which is a reductio ad absurdum. I hope I have sufficiently explained my reasoning for believing as I do.