Feeling the Future: A Meta-analysis of 90 Experiments on the Anomalous Anticipation of Random Future Events

Daryl J. Bem
Cornell University

Patrizio Tressoldi
Università di Padova, Italy

Thomas Rabeyron
Université de Nantes, France and University of Edinburgh, Scotland
and
Michael Duggan
Nottingham Trent University, United Kingdom

Author Note

Daryl J. Bem, Department of Psychology, Cornell University; Patrizio Tressoldi, Dipartimento di Psicologia Generale, Università di Padova, Italy; Thomas Rabeyron, Faculté de Psychologie, Université de Nantes, France and University of Edinburgh, Scotland; Michael Duggan, Nottingham Trent University, United Kingdom.

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. 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.

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

Abstract

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 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 (Jeffries, 1961). The number of potentially unretrieved experiments required to reduce the overall effect size to a trivial value is 547. Several tests demonstrate that the database is not significantly compromised by publication bias, selection bias, or by "*p*-hacking," the selective suppression of findings or statistical analyses that failed to yield statistical significance. An analysis of *p*–curve, the distribution of significant *p* values (Simonsohn, Nelson, & Simmons, 2014a; 2014b) estimates the true effect size of the database to be 0.20, virtually identical to the effect size of Bem's original studies (0.22). We discuss the controversial status of precognition and other anomalous effects collectively known as *psi*.

Keywords: precognition, psi, ESP, retrocausation, retro-priming, parapsychology

Feeling the Future: A Meta-analysis of 90 Experiments on the Anomalous Anticipation of Random Future Events

In 2011, *Journal of Personality and Social Psychology* published an article by Daryl Bem entitled "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." 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 such replications from the very beginning of his research program in 2002, 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 own 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 and 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 recent "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 emotional arousal when these images appear, but the important "presentiment" finding is that an emotional 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, Tressoldi, & Utts, 2012).

Bem's (2011) 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.

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 (Zajonc, 1968; Bornstein, 1989). 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 assess behavioral responses. Even the photographs used in the two kinds of experiments are 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 (2011) 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 (2011) reported a mean effect size (*d*) of 0.22, with a Stouffer *Z* of 6.66, $p = 2.68 \times 10^{-11}$ (Bem, Utts, & Johnson, 2013).

Method[1]

Retrieval and Coding of Experiments

As noted above, the archival summary publication of Bem's experiments appeared in 2011, but he had reported results at annual meetings of the Parapsychological Association as they emerged from his laboratory in 2003, 2005, and 2008, while simultaneously making materials available to those who expressed an interest in trying to replicate the experiments. Reports of the experiments also appeared in the popular media prior to journal publication. As a result, several attempted replications of the experiments were conducted prior to 2011, and are included in our meta-analysis. 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).

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.

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 or not the experiment had been peer-reviewed; e) whether the replication had been conducted before or after the 2011 archival publication of Bem's original experiments; f) whether the investigators appeared to hold favorable or unfavorable initial attitudes and expectations about psi; and g) 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. Eleven experiments were categorized into the third category of experiments that had not been designed to replicate any of Bem's experiments. These included two retroactive priming experiments that assessed non-affective perceptual judgments, three retroactive priming experiments that used the Stroop color-naming task, three experiments that tested the effect of retroactive practice on the speed of shape detection, one experiment that tested the effect of retroactive practice on semantic categorization, and two experiments that tested the effect of retroactive practice on text reading speed.

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.

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, Hedges, Higgins, and Rothstein (2005). Effect sizes (Hedges' *g*) and their standard errors were computed from *t* test values and 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 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, Sanchez-Meca, Marin-Martinez, & Botella, 2006). When I^2 was below 25% we used a fixed-effect model to estimate the parameters; otherwise, we used a random-effects model.[2]

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.

Jeffries (1961) 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 (2011) original experiments (Bem, Utts, & Johnson, 2011; Rouder & Morey, 2011; Wagenmakers et al., 2011).

For purposes of meta-analyis, Rouder and 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, Speckman, Sun, Morey, & Iverson, 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).

The software script in R and the databases are available from Tressoldi (2014).

Results and Discussion

The full database comprises 90 experiments conducted between 2002 and 2013. These originated in 33 different laboratories located in 14 countries and involved 12,406 participants. Fifty-

one (57%) of the experiments had been published in peer reviewed journals or conference proceedings. The complete database of experiments with corresponding effect sizes, standard errors, and category assignments is presented in the Appendix along with a forest plot of the individual effect sizes and their 95% Confidence Intervals.

The primary 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 row of Table 1, the answer is yes: The overall effect size (Hedges' *g*) is 0.09, combined $z = 6.38$, $p = 1.2 \times 10^{-10}$. The Bayesian BF value is 1.4×10^{9}, greatly exceeding the criterion value of 100 that is considered to constitute "decisive evidence" for the experimental hypothesis (Jeffreys, 1961). Moreover, the BF value is robust across a wide range of the scaling factor *r*, ranging from a high value of 4.9×10^{9} when $r = 0.01$ to a low value of 2.0×10^{9} when $r = 1.0$.

A subsidiary question is whether independent investigators can successfully replicate Bem's original experiments. As shown in the second row of Table 1, the answer is again yes: When Bem's experiments are excluded, the combined effect size for attempted replications by other investigators is 0.06, $z = 4.16$, $p = 1.6 \times 10^{-5}$, and the BF value is 3,853, which again greatly exceeds the criterion value of 100 for "decisive evidence."

The next two rows of Table 1 reveal that the mean effect size of exact replications does not differ significantly from that of modified replications (Mean diff = 0.021; 95% CI [-0.04, 0.08]; $z = 0.99$).

And finally, the bottom two rows of the Table demonstrate that replications conducted after the publication of Bem's 2011 article are independently significant and not significantly different from those conducted before its publication (Mean diff = 0.042; 95% CI [.02, 0.10]; $z = 0.37$).

Table 1

Meta-analytic Results for All Experiments and for Independent Replications of Bem's Experiments[3]

	Number of experiments	Number of participants	Effect size (Hedges' *g*)	95% CI[a]	Combined *z* or Bayes Factor ($BF_{H1/H0}$)	I^2	*p*
All Experiments[b]	90	12,406	0.09	[0.06, 0.11]	$z = 6.40$	41.7	1.2×10^{-10}
Bayesian Analysis			0.09	[0.02, 0.15]	BF = 1.4×10^{9}		
Independent Replications of Bem's Experiments[c]	69	10,082	0.06	[0.03, 0.09]	$z = 4.16$	36.1	1.07×10^{-5}
Bayesian Analysis			0.07	[0.01, 0.14]	BF = 3,853		
Exact Replications	31	2,082	0.08	[0.02, 0.13]	$z = 2.78$	33.0	2.7×10^{-3}
Modified Replications	38	8,000	0.06	[0.02, 0.09]	$z = 3.15$	37.7	8.1×10^{-4}
Pre- 2011 Replications	30	2,193	0.09	[0.04, 0.15]	$z = 3.20$	39.5	0.7×10^{-3}
Post-2011 Replications	39	7,899	0.05	[0.02, 0.08]	$z = 2.95$	31.6	1.6×10^{-3}

[a] In a Bayesian analysis, the analogue to CI is referred to as the "credible intervals of the posterior distributions," abbreviated as CrI.

[b] Assuming a null ES of .01 and a variance of 0.0005 (the observed variance in the random-effects model), the statistical power of this meta-analysis is 0.95 (Hedges and Pigott, 2001).

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

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

Table 2
Meta-analytic Results as a Function of Protocol and Experiment Type[3]

Experiment Type	Number of experiments	Number of participants	Effect size	95% CI	I^2	Combined *z*	*p* (One-Tailed)
Fast-Thinking Protocols							
Precognitive Detection of Reinforcement	14	863	0.14[a]	[0.08, 0.21]	19.0	4.22	1.2×10^{-5}
Precognitive Avoidance of Negative Stimuli	8	3,120	0.09	[0.03, 0.14]	50.5	3.10	.002
Retroactive Priming	15	1,154	0.11	[0.03, 0.21]	42.0	2.85	.003
Retroactive Habituation	20	1,780	0.08[a]	[0.04, 0.13]	24.6	3.50	2.3×10^{-4}
Retroactive Practice	4	780	0.12	[0.04, 0.21]	25.5	2.82	.002
All Fast-thinking Experiments	**61**	**7,697**	**0.11**	**[0.08, 0.14]**	**31.6**	**7.11**	**5.8×10^{-13}**
Slow-Thinking Protocols							
Retroactive Facilitation of Practice on Recall	27	4,601	0.04	[-0.01, 0.09]	38.3	1.66	.10
Retroactive Facilitation of Practice on Text Reading Speed	2	108	-0.10	[-0.40, 0.20]	61.0	-0.65	.51
All Slow-thinking Experiments	**29**	**4,709**	**0.03**	**[-0.01, 0.08]**	**39.7**	**1.38**	**.16**

[a] Fixed-effect model

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, the slow-thinking experiments achieved an overall effect size of only 0.04, 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 used modified procedures. The 15 exact replications of that protocol yielded an overall effect size comparable to that of the fast-thinking experiments (0.08), but the 12 modified replications yielded a null effect size (-0.00). For example, Galak, LeBoeuf, Nelson, and Simmons (2012) used their own software to conduct 7 of their 11 modified replications. They ran 3,289 sessions of which 2,845 (86.5%) were unsupervised online sessions that bypassed the controlled conditions of the laboratory. These 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.

Nevertheless, we still believe that it is the fast/slow variable itself that is primarily responsible for the poorer 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 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 3 experiments that were not designed to be replications of Bem's original protocol, then 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, Wetzels, Borsboom, van der Maas, and Kievit (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). This is particularly true 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, $z = 4.74$, $p = 1.07 \times 10^{-6}$.

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 these experimenter variables can have pervasive effects on experimental outcomes with both human and animal subjects (Rosenthal & Rubin, 1978).

Such experimenter effects have also been found in psi research. In an extended psi study specifically designed to investigate experimenter effects, psi proponent Marilyn Schlitz and well-known psi skeptic Richard Wiseman jointly ran three psi experiments in which both investigators used the same procedures and drew participants from the same pool (Schlitz, Wiseman, Radin, & Watt., 2005; Wiseman & Schlitz, 1997, 1999). Schlitz obtained significant positive results in two of the three experiments, but Wiseman obtained null results in all three. (Wiseman also contributed a failed replication to our database [Ritchie, Wiseman, & French, 2012)].

We find similar evidence for experimenter effects in our own database: Considering only exact replications and excluding Bem's own experiments, we find that investigators who held favorable attitudes toward psi obtained higher effect sizes than did those who held unfavorable attitudes: $t(29) = 2.13$, 1-tailed $p = .028$. The judgment of which group's data is

more trustworthy or compelling will undoubtedly be influenced by each reader's own prior attitudes toward psi, but the data themselves are silent regarding such judgments.

Finally, exact replications 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, Nelson, & Simonsohn, 2011). By defining an exact replication in our meta-analysis as one that used Bem's experimental instructions and software, we ensure that the experimental parameters, the stimuli, and the data analyses are all specified ahead of time. In other words, an exact replication is a publicly available, pre-specified protocol that thereby provides many of the same safeguards against false-positive results that are provided by the preregistration of planned experiments.

Despite the merits of exact replications, however, they cannot uncover artifacts in the original protocol that may have produced 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, we saw above that the substitution of mild, non-explicit erotic stimuli may have been responsible for the one failed replication of the erotic stimulus detection experiment reported by Wagenmakers et al. (2012).

To test the generalizability of results in psychological experiments, it has recently been suggested that stimuli should be treat statistically as a random factor the same way we currently treat participants (Judd, Westfall, & Kenny, 2012). If widely adopted, that would

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 across stimuli used in exact replications of Bem's experimental protocols was achieved. In those involving erotic photographs different stimulus sets were used for men and women and all participants were 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.

It is therefore reassuring that both exact and modified replications of Bem's experiments produce comparable, statistically significant results (Table 1), implying 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.

File-Drawer Effects: Missing Studies and *P*-Hacking

Because successful studies are more likely to be published than unsuccessful studies (aka 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.

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 six 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 is 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 possibility that unpublished or unretrieved studies are likely to have mean non-zero effects 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.)

A more conservative 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 547 studies. That is, there would have to 547 studies missing from our database with a mean effect size of .001 to reduce the overall effect size of our database to .01.

The Correlation between Study Size and Effect Size. Another index of publication or retrieval bias is the correlation between the size of a study and its effect size. 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. 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). For our database, Kendall's tau is actually slightly positive: tau = 0.10; z = 1.40; 2-tailed p = 0.15, implying that our database is not seriously biased by a file-drawer effect.

Trim and Fill. An elaborate extension of the correlation between study size and effect size is Duval and Tweedie's (2000) Trim-and-Fill method. It is currently the most common approach to estimating the number of missing studies in a meta-analysis (Simonsohn, Nelson and Simmons, 2014b) and is based on an analysis of 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.

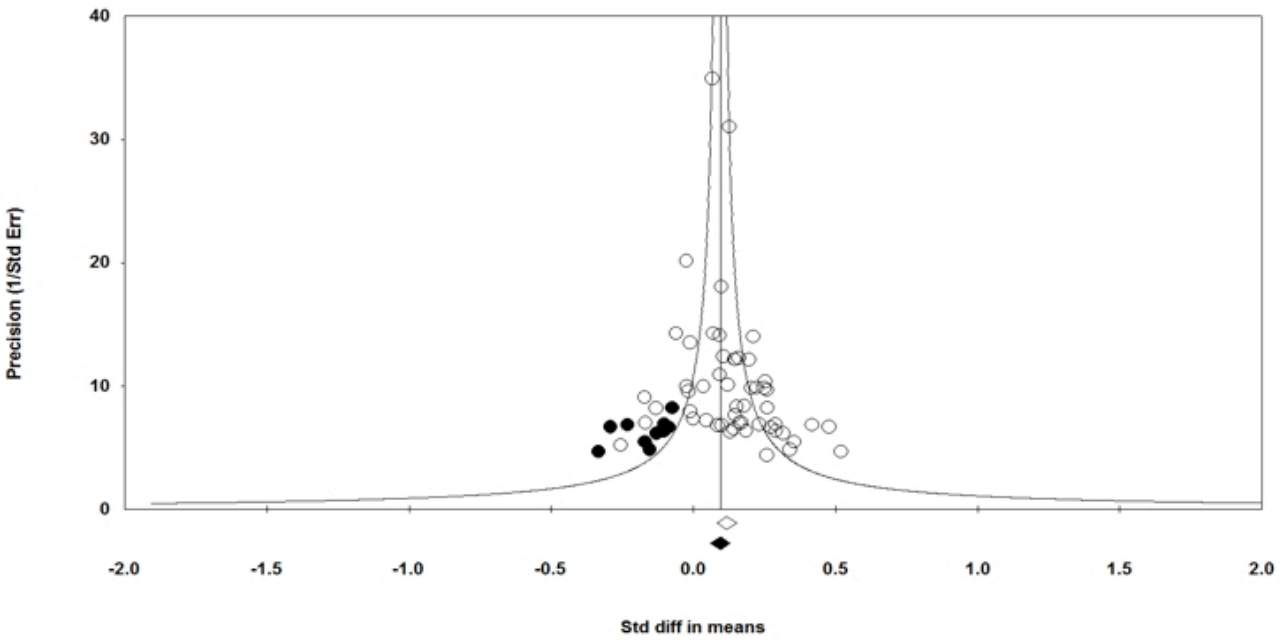

Figure 1. Funnel Plot with the estimated missing studies under a random-effects model.

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 current database.

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 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 true 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 the present database after it has been modified by the trim-and-fill procedure. The unfilled circles identify the actual experiments in the meta-analysis; the black circles identify the imputed missing experiments. The unfilled diamond under the horizontal axis marks the original observed effect size; the black diamond marks the corrected estimate of the effect size. The plot suggests that there are only nine missing experiments, implying again that the file-drawer effect is not a serious problem for our meta-analysis.

The trim-and-fill method has recently been criticized for presuming that publication bias is driven by effect-size considerations even though it is more realistic to suppose that it is driven by the sacred $p = .05$ significance level (Simonsohn, Nelson and Simmons, 2014a; 2014b). These same critics have also demonstrated empirically that trim and fill is inadequate for estimating the true effect size present in the database. In its place, they and other authors

(van Assen, van Aert, & Wicherts, 2014) have recently proposed a quite different approach called, *p*-curve analysis.

***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….If an effect is not real, then 5% of *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, under conditions of no effect …there will be as many *p* values between .04 and .05 as between .00 and .01, and *p*-curve's expected shape is *uniform*.…If an effect exists, 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, p. 666-667).

In their version of *p*-curve analysis, Van Assen et al. (2014) use an algorithm called *p*-uniform to test the degree to which the observed curve differs from a "no-effect" or uniform distribution. For our database, *p*-uniform indicates that there is a significant effect in our database ($p = .004$) but no evidence that there is any publication bias ($p = .876$).

***P*-Curve, Suppressed Data Analyses and *P*-Hacking.** The concern about missing studies from a meta-analysis is long standing, but critical attention to questionable reporting practices within a study, such as selectively presenting only analyses that produced significant results, is more recent (e.g., Simmons, Nelson, & Simonsohn, 2011). In reporting his original set of experiments, Bem (2011) tried to allay such concerns by presenting multiple analyses of each experiment, demonstrating in each case that they all arrived at the same statistical

conclusions. A subsequent Bayesian analysis of the experiments also reaffirmed those conclusions (Bem, Utts, & Johnson, 2013).

Simmons, Nelson, & Simonsohn (2011), have coined the term "*p*-hacking" to describe questionable practices of selective data presentation that will cause the statistical results to meet the coveted $p < .05$ threshold. These same authors have proposed an examination of *p*-curve that can evaluate whether *p*-hacking has compromised the analysis. Specifically,

> 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).

In our database, 19% of the studies reported results that were statistically significant at the .05 level. The *p*-curve distribution for those studies is displayed in Figure 2 and analyzed for skewness in Table 3.

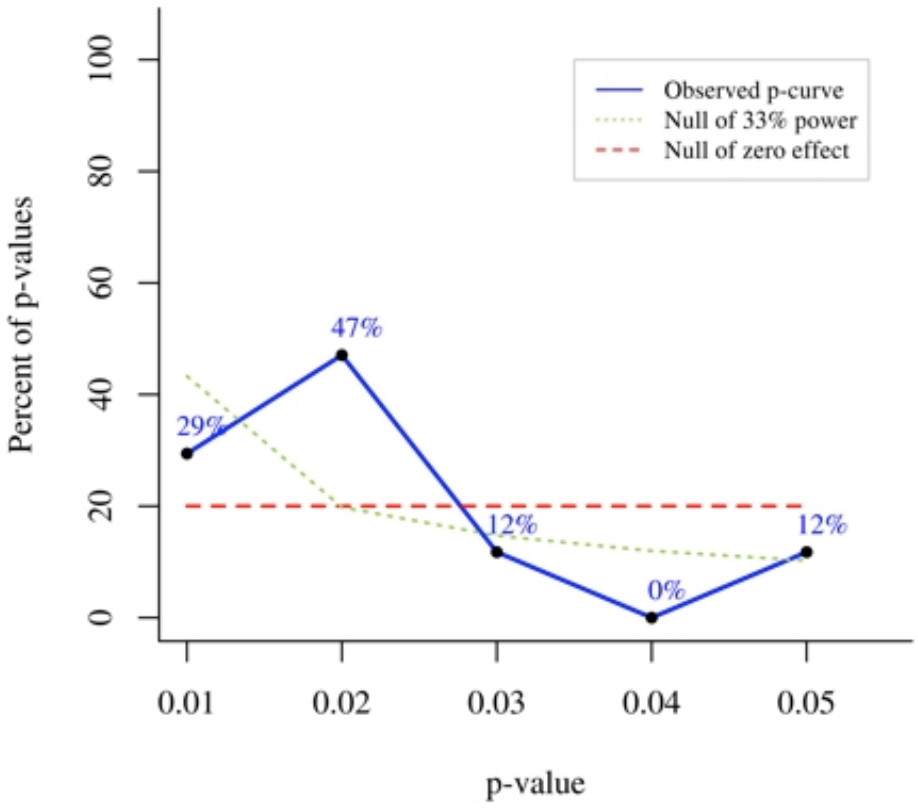

Figure 2. Distribution of the significant *p* values across experiments in the meta-analysis.

Table 3
Skewness Tests on the Distribution of Significant p *Values across Experiments in the Database.*

Statistical Inference	**χ^2 Test**
Studies contain evidential value (*Right-skewed*)	$\chi^2(34) = 47.7$ $p = .059$
Studies lack evidential value (*Flatter than 33%*)	$\chi^2(34) = 34.2$ $p = .45$
Studies lack evidential value and were intensely *p*-hacked (*Left-skewed*)	$\chi^2(34) = 20.1$ $p = .97$

As shown in the first row of Table 3, the *p*-curve falls just short of being significantly right-skewed ($p = .059$). When this is the case, Simonsohn et al. (2014a) propose testing whether it is flatter than one would expect if studies were powered at 33%. As shown in the middle row of the Table, the curve is not flatter than the proposed 33% test. Finally, the bottom row shows that the *p*-curve is clearly not left-skewed. We conclude that *p*-hacking has not significantly distorted our meta-analytic results.

P-Curve and the "True" Effect Size

The *p*-uniform estimate of our database's "true" effect size is 0.11, which is statistically identical to the estimate of 0.09 based on the full database that we report in Table 1. But one of the counter-intuitive 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 in a meta-analysis actually leads to an *underestimate* of the true effect size in the database (Simonsohn et al., 2014b). And, indeed, the *p*-curve estimate of the "true" effect size based on the algorithm published by Simonsohn et al. is 0.20, more then twice the size of these other two estimates. This larger estimate is actually closer to the mean effect size of Bem's (2011) original experiments (0.22) and the effect size of the presentiment experiments (0.21) (Mossbridge et al., 2012).

The discrepancy between the two estimates based on *p*-curve analysis probably derives from the fact that our database is heterogeneous—as indicated by the many I^2 values > 25% and our consequent use of the random-effects model for virtually all our analyses. According to van Assen et al. (2014), the population effect size in their *p*-uniform algorithm "is taken to be fixed rather than heterogeneous (p 4)." In contrast, Simonsohn et al. (2014b) state that "…the accuracy of *p*-curve 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 implies that the higher estimate of .20 for the true effect size in our database is the correct one.

In summary, we find no evidence of publication or selection bias in our database using five different analyses: Orwin's Fail-Safe *N*, the correlation between study size and effect size, Trim-and-Fill analysis of the funnel plot, and two versions of *p*-curve analysis. The *p*-curve analysis also permits us to conclude that questionable reporting practices known as *p*-hacking have not compromised the meta-analysis.

General Discussion

Psi is a controversial subject, and most academic psychologists do not believe that psi phenomena are likely to exist. A survey of 1,100 college professors in the United States revealed that psychologists were much more skeptical about the existence of psi than respondents in the humanities, the social sciences, or the physical sciences, including physics (Wagner & Monnet, 1979). Thus the publication of an article that claims evidence for psi in a mainstream psychological journal should not be taken to imply that the reviewers and editors necessarily agree with the authors' conclusions. This was made explicit in the "Editorial Comment" accompanying Bem's (2011) original article in the *Journal of Personality and Social Psychology*:

> To some of our readers it may be both surprising and disconcerting that we have decided to publish Bem's article.…We openly admit that the reported findings conflict with our own beliefs about causality and that we find them extremely puzzling. Yet, as editors we were guided by the conviction that this paper—as strange as the findings may be—should be evaluated just as any other manuscript on the basis of rigorous peer review (Judd & Gawronski, 2011, p. 406).

There are valid 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 still-sacred $p = .05$ threshold is a constant reminder that it is far more sinful 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 widespread article of faith, see Fiedler, Kutzner, & Krueger, 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, compared with fewer than 4% of all other respondents. 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 a Bayesian critique of Bem's (2011) experiments by Wagenmakers, Wetzels, Borsboom, & van der Maas (2011). Although Wagenmakers et al. did not explicitly claim psi to be impossible, they came very close by setting their prior odds at 10^{20} against the psi hypothesis. The Bayes Factor for our full database is approximately 10^{9} *in favor* of the psi hypothesis (Table 1), 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, Utts, & Johnson, 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 some of them after the fact, (e.g., *p*-curve analysis) but not others (e.g., preregistration). 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 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 (The phenomena of electricity and magnetism were explored by the ancient Greeks as early as 600 BC, but remained without theoretical explanation until Maxwell proposed his field equations in 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 (Radin, 2006).

In pursuit of this possibility, 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, 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 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 the new 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 say the same thing about psi.

References

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

American Psychological Association Publication and Communication Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: why do we need them? What might they be? *American Psychologist, 63*, 839–851.

*Barušs, I. and Rabier, V. (2013). Failure to Replicate Retrocausal Recall.

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

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

*Batthyany, A. (2010). Retrocausal Habituation and Induction of Boredom: A Successful Replication of Bem (2010; Studies 5 and 7) (November 27, 2010). Available at SSRN: http://ssrn.com/abstract=1715954.

*Batthyany, A., Kranza, G.S. and Erber, A. (2009). Moderating factors in precognitive habituation: the roles of situational vigilance, emotional reactivity, and affect regulation. *Journal of Society for Psychical Research*, *73*, 65-82.

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

Bayarri, M. J., & Garcia-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika, 94*, 135 – 152.

Bem, D. J. (2003, August 2-4). Precognitive habituation; Replicable evidence for a process of anomalous cognition. Paper presented at the Parapsychology Association 46th Annual Convention, Vancouver, Canada.

Bem, D. J. (2005, August 11-15). Precognitive aversion. Paper presented at the Parapsychology Association 48th Annual Convention, Petaluma, CA.

Bem, D. J. (2008, August 13-17). 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.

*Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407–425. doi:10.1037/a0021524

*Bem, D. J. (2012). An additional replication of the "precognitive detection of erotic stimuli" experiment. Personal communication.

Bem, D. J., Utts, J., & Johnson, W.O. (2011). Reply to "Must psychologists change the way they analyze their data?" *Journal of Personality and Social Psychology, 101*, 716-719.

*Bierman, D. (2011). Anomalous Switching of the Bi-Stable Percept of a Necker Cube: A Preliminary Study. *Journal of Scientific Exploration, 25*, 4, 721–733.

Bierman, D. J., & Radin, D. I. (1997). Anomalous anticipatory response on randomized future conditions. *Perceptual and Motor Skills, 84*, 689-690.

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

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

Borenstein M., Hedges L. V., Higgins J. P. T., & Rothstein H. R. (2009). *Introduction to meta-analysis*. Wiley: Chichester.

Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. (2005). *Comprehensive meta-analysis* (Version 2). Englewood, NJ: Biostat.

Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1,* 97-111.

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

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

Carpenter, J. C. (2012). *First sight: ESP and parapsychology in everyday life*. Lanham, MD: Rowman & Littlefield.

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

Dijksterhuis, A., & Smith, P. K. (2002). Affective habituation: Subliminal exposure to extreme stimuli decreases their extremity. *Emotion, 2*, 203–214.

DiMaggio, C. (2013). Bayesian Analysis for Epidemiologists Part IV: Meta-Analysis." Injury Control and Epidemiology Pages at Columbia (ICEPaC). Columbia University. Accessed 9 February 2013. <http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/styled-4/styled-11/code-9.

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

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255-278.

Evans, J. S. B. and Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspectives on Psychological Science*, *8*: 223-241. DOI: 10.1177/1745691612460685

Ferguson, M. J. & Zayas, V. (2009). Nonconscious evaluation. *Current Directions in Psychological Science, 18,* 362-366.

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

Fiedler, K., Kutzner, F., and Krueger, J. I. (2012). The long way from α–error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*: 661-669. doi: 10.1177/1745691612462587.

*Fontana, G., Polikarpov, V, and Yankelevich, A. (2012). Experiments on anomalous retroactive influences in the context of the theory of growing block universe. Downloaded from: http://www.chronos.msu.ru/EREPORTS/polikarpov_EXPERIMENTS_OF_D.BEM.pdf.

*Franklin, M.S., and Schooler, J. W. (2013). 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, June 8, 2013, Dearborn, Michigan.

*Galak, Jeff, LeBoeuf, Robyn A., Nelson, Leif D. and Simmons, Joseph P. (2012). Correcting the Past: Failures to Replicate Psi. *Journal of Personality and Social Psychology*, *103*(6), 933-948. doi: 10.1037/a0029709.

Garton, R. (2010). Precognitive priming and sequential effects in visual word recognition. Master Thesis, Macquarie University, Australia.

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

*Hadlaczky G. and Westerlund , J. (2005). 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.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6*, 203–217.

*Hitchman, G.M., Roe, C.A. & Sherwood, S.J. (2012). A re-examination of non-intentional precognition with openness to experience, creativity, psi beliefs and luck beliefs as predictors of success. *Journal of Parapsychology*, *76*,1,109-145.

*Hitchman, G.M., Roe, C.A. & Sherwood, S.J. (2012). The influence of latent inhibition on performance at a non-intentional precognition task. Proceeding of the 55th PA Conference.

*Hitchman, G.M. (2012). Testing the Psi mediated instrumental response theory using an implicit Psi task. Doctoral Thesis, University of Northampton, England.

Honorton, C., & Ferrari, D. C. (1989). "Future telling": A meta-analysis of forced-choice precognition experiments, 1935-1987. *Journal of Parapsychology, 53*, 281–308.

Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F., and Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index?. *Psychological Methods, 11*,193–206.

Judd, C. M., & Gawronski, B. (2011) Editorial comment. *Journal of Personality and Social Psychology, 100*, 406.

Judd, C. M., Westfall, J., & Denny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1) 54-69.

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

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

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

Lang, P. J., & Greenwald, M. K. (1993). *International affective picture system standardization procedure and results for affective judgments.* Gainesville, FL: University of Florida Center for Research in Psychophysiology.

*Luke, D. P., & Morin, S. (2009). 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.

*Luke, D. P., Delanoy, D. & Sherwood, S. J. (2008). Psi may look like luck: Perceived luckiness and beliefs about luck in relation to precognition. *Journal of Society for Psychical Research*, 72, 193–207.

*Luke, D. P., Roe, C. A., & Davison, J. (2008). Testing for forced-choice precognition using a hidden task: Two replications. *Journal of Parapsychology*, 72, 133–154.

*Macadam reported in Wiseman, R, and Watts, C. (submitted). Back to the future: A pre-registry of precognition studies.

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

*Maier M.A., Büchner, V.L. Kuhbandner, C., Pflitsch, M., Fernández-Capo, M., & Gámiz-Sanfeliu, M. (2014). Feeling the future again: Retroactive avoidance of negative stimuli. *Journal of Consciousness Studies. 21*(9-10), 121-152.

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

Morey, R. D. & Rouder, J. N. (2014). BayesFactor: Computation of Bayes factors for common designs. R Package version 0.9.9.

*Morris, B. (2004). Precognitive habituation. Daryl Bem's personal communication.

Mossbridge J, Tressoldi P and Utts J. (2012). Predictive physiological anticipation preceding seemingly unpredictable stimuli: a meta-analysis. *Frontiers of Psychology 3*:390. doi:10.3389/fpsyg.2012.00390.

*Moulton, S.T. (2011). PSI and psychology: the recent debate. Personal communication.

O'Donohue, W. T., & Fisher, J. E. (Eds.). (2009). *General principles and empirically supported techniques of cognitive behavior therapy*. Wiley.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics. 8*, 157–159.

*Parker, A., & Sjödén, B. (2010). Do some of us habituate to future emotional events? *Journal of Parapsychology*, *74*, 99–115.

*Pedersen, J. C., Shepardson, S. K., Lemka, Z. R., & Harton, H. C. (2012). 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.

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

Popa, I.L. & Batthyany, A. (2012). Retrocausal Habituation: A Study on Time-Reversal Effects in the Human Information Processing. Paper presented at the Cognitive Science Conference, Bratislava.

*Rabeyron, T.(2014). Retro-priming, priming and double testing : psi and replication in a test-retest design. *Front. Hum. Neurosci*.8:154. doi:10.3389/fnhum.2014.00154

*Rabeyron, T., & Watt, C. (2009). Paranormal experiences, mental health and mental boundaries, and psi. *Personality and Individual Differences* doi:10.1016/j.paid.2009.11.029.

Radin, D. I. (2006). *Entangled minds: Extrasensory experiences in a quantum reality*. New York, NY: Paraview Pocket Books.

*Ritchie S. J., Wiseman R., & French C. C. (2012) Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PLoS One*, *7*(3): e33423. doi:10.1371/journal.pone.0033423.

*Robinson, E. (2011). Not Feeling the Future: A Failed Replication of Retroactive Facilitation of Memory Recall. *Journal of the Society for Psychical Research*, *75*,3, 142-147.

*Roe, C., Grierson, S. and Lomas, A. (2012). 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.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, 775-777.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Newbury Park, California: Sage.

Rosenthal, R., & Rubin, D. B. (1978) Interpersonal expectancy effects: the first 345 studies. *The Behavioral and Brain Sciences, 3,* 377-415.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta analysis: prevention, assessment and adjustments*. West Sussex, England: Wiley.

Rouder, J. N., and Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*. *18*, 682-689. doi: 10.3758/s13423-011-0088-7.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2) 225-237. doi: 10.3758/PBR.16.2.

*Savitsky, K. (2003). 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.

*Savva, L., & French, C. C. (2002). Is there time-reversed interference in Stroop- based tasks. The Parapsychological Association 45th Annual Convention, Proceedings of the Presented Papers, 194-205.

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

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

Schlitz, M., Wiseman, R., Radin, D., & Watt, C. (2005, August). *Of two minds: Skeptic-proponent collaboration within parapsychology.* Paper presented at the meeting of the Parapsychological Association, Petaluma, CA.

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

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

*Simmonds-Moore, C.A. (2013). Exploring the Relationship between the synaesthesias and anomalous experiences. Unpublished final report to the Bial foundation

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366. doi: 10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). *P*-Curve: A Key to the File-Drawer. *Journal of Experimental Psychology*. *General*, *143*, 534-547. doi: 10.1037/a0033242

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). *p*-Curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*: 666-681. DOI: 10.1177/1745691614553988

Snodgrass, S. (September 27, 2011). Examining retroactive facilitation of recall: an adapted replication of Bem (2011, Study 9) and Galak and Nelson (2010). Available at SSRN: http://ssrn.com/abstract=1935942 or http://dx.doi.org/10.2139/ssrn.1935942

*Starkie, A. (2009). Retroactive habituation: Exploring the time reversed amygdalar response to pictures of facial affect. Dissertation presented at the Liverpool Hope University. Available online at: http://www.slashdocs.com/nvmyq/retroactive-habituation-exploring-the-time-reversed-amygdalar-response-to-pictures-of-facial-affect.html

Storm, L., Tressoldi, P. E., & Di Risio, L. (2013) Meta-analysis of ESP studies, 1987-2010: Assessing the success of the forced-choice design in parapsychology. *Journal of Parapsychology, 76,* 243-273.

*Subbotsky, E. (2013). Sensing the Future: Reversed Causality or a Non-standard Observer Effect?. *The Open Psychology Journal, 6,* 81-93.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals, *Psychological Monographs, 2*, 8.

*Traxler, M.J., Foss, D.J., Podali, R, Zirnstein M. (2012). Feeling the past: The absence of experimental evidence for anomalous retroactive influences on text processing. *Memory and Cognition*, 1-7. DOI 10.3758/s13421-012-0232-2.

Tressoldi, P. (2014). Meta-analysis Implicit Behavioral Anticipation. figshare.http://dx.doi.org/10.6084/m9.figshare.903716

*Tressoldi, P. E., Masserdotti, F., & Marana C. (2012). 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 from http://www.PsychFileDrawer.org.

*Tressoldi, P.E., Zanette, S. (2012). Feeling the future: an exact replication of the Retroactive Facilitation of Recall II and Precognitive Positive Detection experiments with Italian participants uploaded on http://www.PsychFileDrawer.org.

van Assen, M. L. M., van Aert, R. C. M., & Wicherts, J. M. (2014) Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, doi: 10.1037/met0000025

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

*Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632-638.

Wagenmakers, E.J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology, 100*, 426-432. doi:10.1037/a0022790.

Wagner, M. W., & Monnet, M. (1979). Attitudes of college professors toward extra-sensory perception. *Zetetic Scholar, 5*, 7–17.*Watt, C., & Rabeyron, T. (2010). Retro-priming and double testing. In: Proceedings of the 53rd Annual Convention of the Parapsychological Society. (pp. 6). Parapsychological Association.

*Watt, C. & Nagtegaal, M. (1999). Luck in action? belief in good luck, Psi-mediated instrumental response, and games of chance. *Journal of Parapsychology, 63,* xxx-xxx.

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

Wiseman, R., & Schlitz, M. (1999, August). Replication of experimenter effect and the remote detection of staring. Paper presented at the 43rd Annual Convention of the Parapsychological Association, Palo Alto, California.

*Wiseman, R, and Watts, C. (submitted). Back to the future: A pre-registry of precognition studies.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*, 1–29.

*Zangari, W. (2006). Replication of the retro-habituation effect. Personal communication.

Footnotes

[1]The methodology and reporting of results comply with the Meta-Analysis Reporting Standards (APA, 2008).

[2] The effect size estimator, Hedges' *g*, is similar to the more familiar Cohen's *d*, but pools studies using *n* - 1 for each sample instead of *n*. This provides a better estimate for smaller sample sizes. A fixed-effect model assumes that all the studies using a particular protocol (e.g., the set of retroactive priming experiments) 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. The heterogeneity statistic I^2 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. (Borenstein, Hedges, Higgins, & Rothstein, 2009).

[3]Modified versions of Tables 1 and 2 are reproduced in the Appendix as supplementary Tables S1 and S2, respectively, with the between-studies variance measure τ^2 substituting for the heterogeneity statistic I^2, and *p* levels.

Table A1

Experiments in the meta-analysis, N, task type, effect size, standard error, peer-review and replication classifications

Study	Year	Task	N	ES	SE	Peer Reviewed	Exact Replication
Baruss	2013	word recall	102	-0.014	0.098	No	Yes
Batthyány	2009	priming	120	0.093	0.091	No	Yes
Batthyány & Spajic	2008	habituation	43	0.139	0.151	No	Yes
Batthyány	2010	habituation	70	0.205	0.119	No	No
Batthyàny	2008	priming	50	0.471	0.147	No	Yes
Batthyány, Kranz & Erber	2009	habituation	46	0.268	0.148	Yes	Yes
Bem exp1	2011	reward	100	0.249	0.101	Yes	Yes
Bem exp2	2011	avoidance	150	0.194	0.082	Yes	Yes
Bem exp3	2011	priming	97	0.257	0.102	Yes	Yes
Bem exp4	2011	priming	99	0.202	0.101	Yes	Yes
Bem exp5	2011	habituation	100	0.221	0.100	Yes	Yes
Bem exp6	2011	habituation	150	0.145	0.082	Yes	Yes
Bem exp7	2011	habituation	200	0.092	0.071	Yes	Yes
Bem exp8	2011	word recall	100	0.191	0.100	Yes	Yes
Bem exp9	2011	word recall	50	0.412	0.145	Yes	Yes
Bem 2012	2012	reward	42	0.285	0.155	No	Yes
Bierman	2011	priming	169	0.108	0.077	Yes	na
Bierman & Bijl	2013	retro-practice	67	0.313	0.124	Yes	na
Boer, De R., & Bierman	2006	priming	51	0.411	0.144	Yes	na
Cardena et al.	2009	word recall	38	-0.043	0.159	Yes	No
Fontana et al.	2012	reward	59	0.145	0.129	No	No
Franklin lab	2012	retro-practice	194	0.139	0.072	No	na
Franklin online	2012	retro-practice	416	0.061	0.049	No	na

Galak exp 1	2012	word recall	112	-0.113	0.094	Yes	Yes
Galak exp 2	2012	word recall	158	0.000	0.079	Yes	No
Galak exp 3	2012	word recall	124	0.114	0.090	Yes	No
Galak exp 4	2012	word recall	109	0.168	0.096	Yes	No
Galak exp 5	2012	word recall	211	-0.049	0.069	Yes	No
Galak exp 6	2012	word recall	106	-0.029	0.096	Yes	No
Galak exp 7	2012	word recall	2469	-0.005	0.020	Yes	No
Hadlaczky & Westerlund	2005	habituation	47	0.085	0.144	No	Yes
Hitchman study2	2012	reward	50	-0.050	0.139	No	No
Hitchman study4	2012	reward	52	0.044	0.137	No	No
Hitchman et al	2012	reward	50	0.159	0.140	Yes	No
Hitchman et al B	2012	reward	49	0.228	0.142	Yes	No
Luke & Morin	2009	reward	41	0.182	0.155	No	No
Luke, Delaoy & Sherwood	2008	reward	100	0.249	0.101	Yes	No
Luke, Roe, Davison study 1	2008	reward	25	0.504	0.206	Yes	No
Luke, Roe, Davison study 2	2008	reward	32	0.347	0.178	Yes	No
Macadan	2011	word recall	88	0.026	0.106	No	Yes
Maier study failed 1	2012	avoidance	63	-0.012	0.124	Yes	No
Maier study failed 2	2012	avoidance	406	-0.024	0.050	Yes	No
Maier study1	2012	avoidance	111	0.251	0.096	Yes	No
Maier study2	2012	avoidance	201	0.210	0.071	Yes	No
Maier study3	2012	avoidance	1222	0.068	0.029	Yes	No
Maier study4	2012	avoidance	327	0.100	0.055	Yes	No
Maier failed 3	2013	avoidance	640	0.052	0.040	Yes	No
Milyavsky	2010	word recall	58	-0.012	0.130	No	Yes
Morris	2004	habituation	40	0.313	0.159	No	Yes
Moulton & Kosslyn	2001	habituation		0.178	0.118	No	No

Moulton & Kosslyn	2004	habituation		-0.010	0.074	No	No
Moulton_b	2004	habituation		-0.060	0.070	No	No
Moulton	2004	habituation		0.070	0.070	No	Yes
Moulton	2003	priming		-0.129	0.121	No	Yes
Parker and Sjodén	2010	habituation	20	0.249	0.218	No	Yes
Pedersen et al.	2012	word recall	96	0.186	0.102	No	Yes
Platzer	2012	word recall	98	0.111	0.101	No	No
Popa and Batthyany	2012	habituation	50	-0.142	0.140	No	No
Rabeyron and Watt	2010	priming	155	0.106	0.080	Yes	No
Rabeyron	2010	priming	28	-0.248	0.187	Yes	No
Ritchie et al. Exp1	2012	word recall	50	0.015	0.139	Yes	Yes
Ritchie et al. Exp2	2012	word recall	50	-0.219	0.141	Yes	Yes
Ritchie et al. Exp3	2012	word recall	50	-0.040	0.139	Yes	Yes
Robinson	2011	word recall	50	-0.118	0.140	Yes	No
Roe, Grierson, Lomas 1	2012	priming	47	0.099	0.144	No	Yes
Roe, Grierson, Lomas 1b	2012	word recall	50	0.078	0.139	No	No
Roe, Grierson, Lomas 2	2012	priming	42	-0.096	0.152	No	Yes
Roe, Grierson, Lomas 2b	2012	word recall	50	-0.042	0.139	No	No
Savitsky	2003	habituation	84	0.170	0.109	No	No
Savva & French exp1	2002	priming	40	0.128	0.156	Yes	na
Savva & French exp2	2002	priming	50	0.166	0.140	Yes	na
Savva & French exp3	2002	priming	54	0.000	0.134	Yes	na
Savva et al. phobic	2004	habituation	25	0.329	0.199	Yes	Yes
Savva et al. Study 1	2005	habituation	50	0.284	0.142	Yes	Yes
Savva et al. Study 2	2005	habituation	92	-0.018	0.103	Yes	Yes
Simmonds-Moore	2013	word recall	52	0.049	0.137	No	Yes
Starkie	2009	habituation	50	-0.163	0.140	No	Yes

Subbotsky exp 1	2013	word recall	75	0.279	0.117	Yes	Yes
Subbotsky exp 2	2013	word recall	25	0.292	0.198	Yes	Yes
Subbotsky exp 3	2013	word recall	26	-0.399	0.198	Yes	Yes
Traxler et al. Exp1a	2012	text speed	48	0.060	0.142	Yes	na
Traxler et al. Exp1b	2012	text speed	60	-0.249	0.129	Yes	na
Tressoldi et al.	2012	priming	100	0.036	0.099	No	Yes
Tressoldi et al. recall	2012	word recall	100	0.221	0.100	No	Yes
Tressoldi, Zanette	2012	reward	103	0.120	0.098	No	Yes
Tressoldi, Zanette	2012	word recall	104	-0.007	0.097	No	Yes
Vernon	2013	retro-practice	102	0.152	0.099	No	na
Wagenmakers et al.	2012	reward	100	-0.022	0.099	Yes	No
Watt & Nagtegaal	2000	reward	60	-0.076	0.128	No	No
Zangari	2006	habituation	52	0.046	0.137	No	Yes

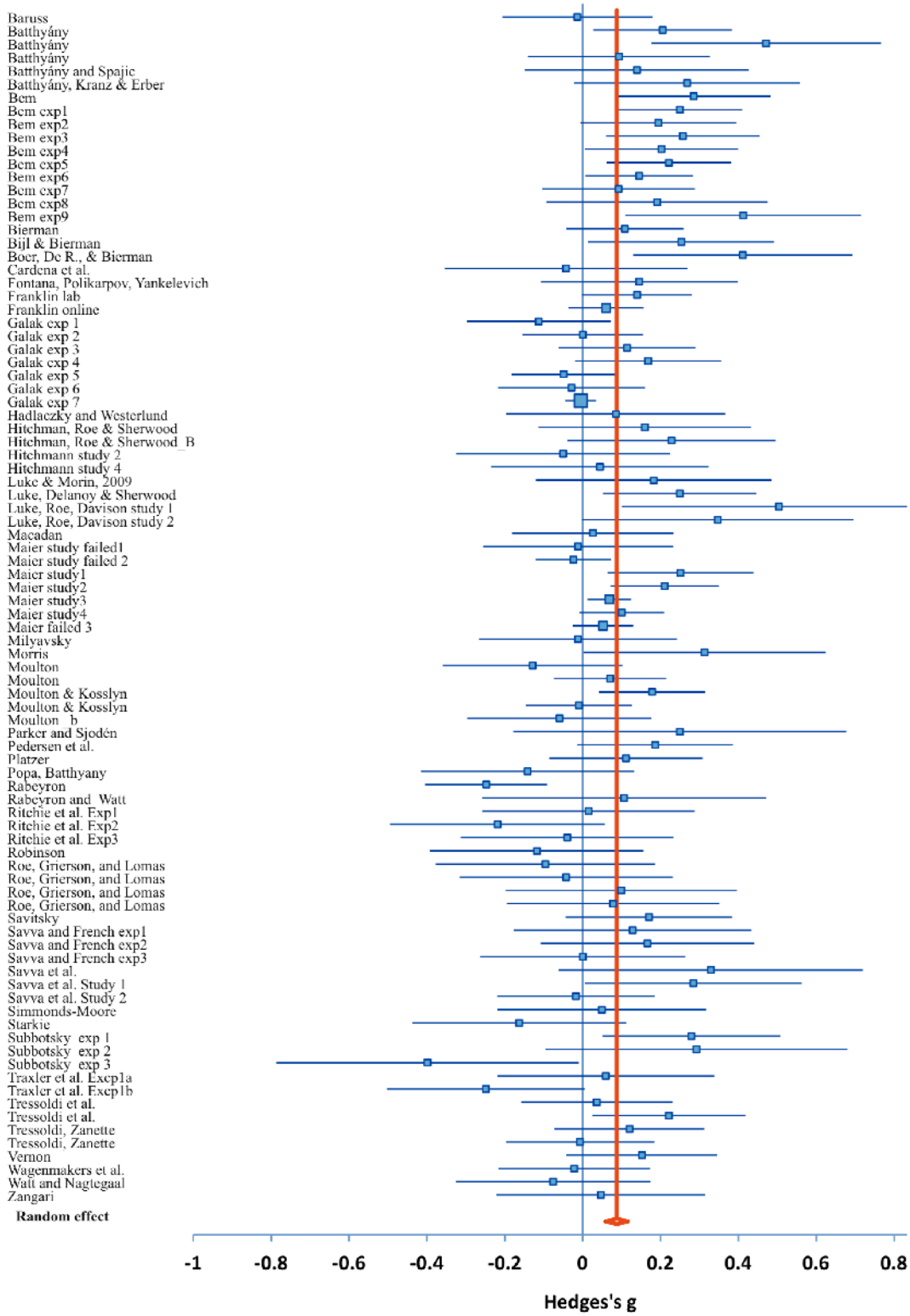

Figure A1. Forest plot of effect sizes. Each blue spot identifies the estimated effect size for that experiment with the corresponding 95% confidence interval. The red vertical line marks the overall effect size based on the random-effects model.

Supplementary Material

Table S1

Meta-analytic Results for All Experiments and for Independent Replications of Bem's Experiments with a column of τ^2 added

	Number of experiments	Number of participants	Effect size (Hedges' *g*)	95% CI[a]	Combined *z* or Bayes Factor ($BF_{H1/H0}$)	τ^2
All Experiments[b]	90	12,406	0.09	[0.06, 0.11]	$z = 6.40$	0.005
Bayesian Analysis			0.09	[0.02, 0.15]	BF = 1.4×10^9	0.028
Independent Replications of Bem's Experiments[c]	69	10,082	0.06	[0.03, 0.09]	$z = 4.16$	0.004
Bayesian Analysis			0.07	[0.01, 0.14]	BF = 3,853	0.035
Exact Replications	31	2,082	0.08	[0.02, 0.13]	$z = 2.78$	0.007
Modified Replications	38	8,000	0.06	[0.02, 0.09]	$z = 3.15$	0.003
Pre- 2011 Replications	30	2,193	0.09	[0.04, 0.15]	$z = 3.20$	0.009
Post-2011 Replications	39	7,899	0.05	[0.02, 0.08]	$z = 2.95$	0.003

[a] In a Bayesian analysis, the analogue to CI is referred to as the "credible intervals of the posterior distributions," abbreviated as CrI.

[b] Assuming a null ES of .01 and a variance of 0.0005 (the observed variance in the random-effects model), the statistical power of this meta-analysis is 0.95 (Hedges and Pigott, 2001).

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

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

Table S2

Meta-analytic Results as a Function of Protocol and Experiment Type with a column of τ^2 added

Experiment Type	Number of experiments	Number of participants	Effect size	95% CI	τ^2
Fast-Thinking Protocols					
Precognitive Detection of Reinforcement	14	863	0.14[a]	[0.08, 0.21]	0.004
Precognitive Avoidance of Negative Stimuli	8	3,120	0.09	[0.03, 0.14]	0.003
Retroactive Priming	15	1,154	0.11	[0.04, 0.20]	0.009
Retroactive Habituation	20	1,780	0.08[a]	[0.04, 0.13]	0.004
Retroactive Practice	4	780	0.12	[0.04, 0.18]	0.002
All Fast-thinking Experiments	**61**	**7,697**	**0.11**	**[0.08, 0.14]**	**0.004**
Slow-Thinking Protocols					
Retroactive Facilitation of Practice on Recall	27	4,601	0.04	[-0.01, 0.09]	0.005
Retroactive Facilitation of Practice on Text Reading Speed	2	108	-0.14	[-0.40, 0.20]	0.029
All Slow-thinking Experiments	**29**	**4,709**	**0.03**	**[-0.01, 0.08]**	**0.005**

[a] Fixed-effect model