Journal of Personality and Social Psychology
2012, Vol. 103, No. 6, 933–948

0022-3514/12/$12.00 DOI: 10.1037/a0029709

Correcting the Past: Failures to Replicate Psi

Jeff Galak
Carnegie Mellon University

Robyn A. LeBoeuf
University of Florida

Leif D. Nelson
University of California, Berkeley

Joseph P. Simmons
University of Pennsylvania

Across 7 experiments (N = 3,289), we replicate the procedure of Experiments 8 and 9 from Bem (2011), which had originally demonstrated retroactive facilitation of recall. We failed to replicate that finding. We further conduct a meta-analysis of all replication attempts of these experiments and find that the average effect size (d = 0.04) is no different from 0. We discuss some reasons for differences between the results in this article and those presented in Bem (2011).

Keywords: psi, precognition, ESP, researcher degrees of freedom, meta-analysis

Recently, Bem (2011) published an extremely thought-provoking article demonstrating the existence of precognition, a "conscious cognitive awareness . . . of a future event that could not otherwise be anticipated through any known inferential process" (p. 407). Through nine experiments, Bem found consistent support for the idea that people have such precognitive abilities. He suggested that these findings present examples of retroactive influence, through which future events influence people's current responses and that more broadly these findings are instances of *psi phenomena*, or "anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms" (Bem, 2011, p. 407).

In his article, Bem (2011) acknowledged that psi is a controversial topic. He reported data suggesting that many, if not most, academic psychologists do not believe that psi phenomena exist. Indeed, the publication of Bem's research met with a wide variety of reactions in the academic and popular media alike, and although some reactions were supportive, many were skeptical (Carey, 2011a; Carey, 2011b; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). In light of the skepticism surrounding psi and in anticipation of the reaction to his article, Bem suggested that psi researchers must conduct tightly controlled experiments that demonstrate psi and "that can be replicated by independent investigators" (Bem, 2011, p. 407). Whereas Bem's article may indeed provide the necessary tightly controlled experiments, the purpose of the current article is to conduct and to synthesize replications by independent investigators.

Psi Phenomena

The precognitive abilities reported by Bem (2011) emerged across a range of tasks. As one example, in Experiment 1, Bem (2011) asked participants to select whether a picture would appear on the left side of the screen or the right side of the screen. Participants' selections were accurate more often than chance would predict when the picture in question was an erotic one (but not a neutral, positive, or negative one), suggesting that people have precognitive abilities to detect where erotic stimuli will appear.

Precognitive abilities also manifested on more complicated tasks. For example, in Experiment 5, participants were asked to choose which of two negatively arousing pictures they liked better. After this choice, the computer randomly selected one of the pictures to serve as the target picture, which then flashed subliminally on the screen from 4 to 10 times. Research on the mere-exposure effect suggests that subliminal exposure to a negative target increases liking of that target (i.e., causes habituation; Kunst-Wilson & Zajonc, 1980). Bem (2011) suggested that if people have precognitive abilities, their current liking of a negative picture would be enhanced by the fact that they will see that picture several times in the future (even though they have no known way of knowing that they will see it). Bem's results supported this prediction: When participants chose between negative picture pairs, they were more likely to prefer the one that would later be selected to be the subliminally presented target.

Perhaps the most straightforward and impressive demonstration of precognition emerged in Bem's (2011) Experiments 8 and 9, which documented "retroactive facilitation of recall" (p. 419). In these studies, participants saw 48 words and then were asked to recall as many of those words as possible. Next, participants were given a chance to practice a randomly chosen subset of the 48 words by, for example, retyping them and recategorizing them. In

This article was published Online First August 27, 2012.

Jeff Galak, Tepper School of Business, Carnegie Mellon University; Robyn A. LeBoeuf, Department of Marketing, University of Florida; Leif D. Nelson, Department of Marketing, University of California, Berkeley; Joseph P. Simmons, Operations and Information Management Department, University of Pennsylvania.

We would like to thank Eduardo Andrade and Justin Kruger for crucial assistance with the project and Gia Nardini for her help in collecting data for this article.

Correspondence concerning this article should be addressed to Jeff Galak, Tepper School of Business, Carnegie Mellon University, 5000 Forbes Avenue, Office 381-D, Pittsburgh, PA 15213. E-mail: jgalak@cmu.edu

a typical memory test, practice would occur before recall, and one would expect recall of the practiced words to be superior to recall of the unpracticed words. In Bem's (2011) experiment, practice occurred after the recall stage, but Bem suggested that the to-be-practiced words might "reach back in time" (Bem, 2011, p. 419) to enhance the recall of those words. Indeed, the to-be-practiced words were more likely to appear in the recalled set of words than were the words that would not be practiced, consistent with the idea that people have a precognitive ability that leads them to be influenced by *future* practice and not just by practice that has already happened. These results emerged even though there was no discernible way for participants to know which words would be practiced.

Replicating Bem (2011)

Bem (2011) called for independent investigators to replicate his procedures. One purpose of this article is to do precisely that. We conducted these experiments with a formally agnostic stance: We were not trying to "prove psi" or "disprove psi," but rather we were trying to offer more data to bring to bear on the phenomenon. That said, we recognize that researchers' own beliefs can influence the results that they obtain, and so we tried to remove any subjectivity and experimenter influence from our experiments. As described in the Method section, we used Bem's exact procedures and materials whenever we could, and we used computers to standardize the delivery of the instructions and materials. We also predetermined our intended samples (e.g., "a minimum of 100 participants"), and always formally stopped the experiment before looking at any results. We used the same data analytic strategies that Bem used, and we also heeded the advice of Wagenmakers et al. (2011) to use additional analyses, in particular Bayesian *t* tests (described in more detail later).

Altogether, we ran seven experiments with seven different samples, examining over 3,000 participants. We focused our replication attempts on the retroactive facilitation of recall findings described above: Four experiments replicated the procedures of Bem's (2011) Experiment 8, and three experiments replicated the procedures of Bem's (2011) Experiment 9. We chose these findings in particular because the other findings reported in Bem (2011) hinge on nuanced affective responses, such as arousal to erotic images or a preference for avoiding negative images. As Bem (2011) reported, one difficulty with such experiments is that finding the appropriate stimuli can be difficult (e.g., people can foresee erotic images only if they are sufficiently erotic, and men and women require different erotic stimuli and different negative stimuli). Thus, the findings involving affective responses seem to be sensitive to subtle variation in the intensity and character of the stimuli. Not only is extensive pretesting required to find the right stimuli but this need for appropriate stimuli makes it easy to dismiss any null findings as due to the use of inappropriate stimuli.

In the retroactive facilitation of recall studies, on the other hand, people are simply shown a list of words and are then asked to freely recall as many as possible. Participants are then randomly assigned to practice half of the words, with precognition being observed if people recall more of the words that they subsequently practice than words that they subsequently do not practice. In comparison to the other studies reported by Bem (2011), practicing and remembering words was relatively straightforward for us to replicate without concerns about the stimuli insufficiently matching the parameters suggested in the original article. In fact, as noted below, we used the exact stimuli used by Bem (2011) in four of our experiments.

In addition to replicating Bem's (2011) retroactive facilitation of recall studies, another goal of this article was to conduct a meta-analysis of all attempts to replicate these particular studies. We should note that other meta-analyses of psi phenomena have been conducted, but they are not of direct relevance to our conclusions because they do not examine the retroactive facilitation of recall paradigm. Nevertheless, they are worth consideration. Milton (1997) found evidence for a wide range of parapsychological phenomena but warned that the vast majority of experiments did not predefine their outcome measure and therefore should be greatly discounted. Dunne and Jahn (2003) concluded that evidence for remote perception is relatively weak and, from a meta-analytic point of view, is nonexistent. Storm, Tressoldi, and Di Risio (2010) concluded that evidence for psychic communication (i.e., telepathy) does, in fact, persist across a variety of testing conditions. Finally, Tressoldi (2011) conducted a meta-analysis of these three published meta-analyses and two additional unpublished analyses and concluded that, using a frequentist data analytic approach, there is substantial evidence for psi, but using Bayesian analyses, there is mixed evidence for psi. As noted, however, these meta-analyses do not include Bem's (2011) tightly controlled psi experiments. Thus, one of the central goals of this article, aside from directly attempting to replicate Bem's retroactive facilitation of recall experiments, is to conduct a new meta-analysis that includes both our new empirical findings and all other attempted replications of these particular experiments.

Method

Below, we briefly review the basic methodology of our replication attempts. We then provide the relevant details about the specifics of data collection in each experiment. Because the seven experiments that we conducted were highly similar to each other, we present the methods of all seven experiments before turning to their results. This report adheres to the requirements proposed by Simmons, Nelson, and Simonsohn (2011).

All instructions and manipulations were presented through a computer interface. As in Bem (2011), participants first read and agreed to a consent form mentioning that the experimenter was investigating extrasensory perception (ESP) and then read a brief introductory statement almost identical to the one used by Bem (2011):

> This experiment tests for ESP (extra sensory perception) by administering several tasks involving common everyday words. The experiment takes about 15 minutes to complete. The program will give you specific instructions as you go. At the end of the session, the computer will explain to you how this procedure tests for ESP.

When participants had finished reading the statement (after a forced time delay of 7 s to better ensure that participants read the text), they clicked to advance to the next screen.

On the two subsequent screens, participants answered the same stimulus-seeking items that Bem (2011) reported administering. Both items were preceded by, "To what extent is the following statement true of you:" The first item was "I am easily bored," and

the second was "I often enjoy seeing movies I've seen before." Participants responded on 5-point scales anchored at 1 ("*Very Untrue*") and 5 ("*Very True*").

Participants then experienced a 3-min relaxation procedure as described in Bem (2011): They looked at an astronomical photograph while listening to relaxing music. When the 3 min had ended, participants clicked a button to acknowledge that they were ready. Based on the procedure outlined by Bem, they then received these instructions about the task:

> Next, we would like you to look at a list of 48 common nouns one at a time, for 3 seconds. While looking at each word, please visualize the corresponding object. For example, if the word is "house," please imagine a house. When you are ready to begin, please click continue.

Participants in Experiments 1, 2, 6, and 7, who completed the experiments online, were given an additional instruction: "It is absolutely critical that you focus on only this task and do not perform any other tasks (e.g., check e-mail)."

After participants clicked "continue," they were shown the series of words, each for 3 s. We completed our first two experiments and began data collection for our seventh experiment prior to Bem (2011) making his exact materials publically available. Accordingly, we created the lists of words ourselves. In Experiments 1 and 7 we used the same four categories as Bem (2011; food, animals, occupations, and clothes), and for Experiment 2 we created four new categories (kitchen items, electronics, body parts, sports). For the remaining experiments, we used exactly the set of words used by Bem (2011). Appendix A presents the full lists of words for Experiments 1 through 7. Paralleling Bem's procedure, the words were presented in a predetermined random order (the same order for all participants). After all 48 words had been presented, participants were asked to type any words that they recalled. They had as much time as they wanted, and when they were finished, they clicked a button to go to the next stage.

At that point the program, using a pseudorandom number generator, randomly assigned 24 words to be practiced; six words were randomly chosen from each of the four groups of 12 words. Practice unfolded as follows: Replicating Bem's (2011) Experiment 9, participants in our Experiments 4 through 6 were shown and asked to visualize the 24 practice words one at a time for 3 s. Specifically, they were given the following instructions: "You will now be shown 24 of the words you saw earlier, divided into 4 categories: Foods, Animals, Occupations, and Clothing. As you see each word, try to form an image of the thing it refers to (e.g., if the word is *tree*, visualize a tree)." Consistent with Bem's Experiment 8, participants in our Experiments 1, 2, 3, and 7 did not complete this first practice task. Next, all participants in every experiment viewed the list of 24 practice words. On successive screens, they were asked first to click on the six words from a specified category (at which point the words became highlighted) and then to retype those words in six boxes below. Participants could not continue until they correctly clicked on the appropriate six words and typed the six words in the corresponding boxes. They did this for each of the four categories, as in Bem (2011).

Participants in Experiments 1, 2, 6, and 7 (the online experiments) answered one more question:

> It is very important for us to know if you were not paying 100% attention to this study (e.g., checking e-mail, going to the bathroom). You will not be penalized in any way if you did other tasks, and you will be entered into the lottery regardless of how you respond. So please be honest! Did you, at any point during this study, do something else (e.g., check e-mail)?

Participants could check a box corresponding to either "No, I paid 100% attention to the study" or "Yes, I did other things during the study."

Finally, because of the open nature of Experiment 7 (details below), participants in this experiment answered one more question: "Is this your first time taking this experiment (or one similar to it)?" Participants could check a box corresponding to either "No, I've taken this experiment before" or "Yes, I've never taken this experiment before."

For each experiment, we specify how we determined sample sizes, but it is worth an additional mention that in all cases we did not download any of the data prior to terminating any experiment. For all cases, we sought at least 100 participants to mirror the number of participants in Bem's (2011) Experiment 8. In the cases where we set a target of greater than 100, this was largely done to make sure that the samples were large enough to be considered a fair replication attempt.

Experiment 1

Participants ($n = 112$; 88 female, 23 male, 1 unknown; median age = 38) were recruited from an online panel to complete the experiment for a chance to win a $100 gift card. All participants were registered members of the website *consumerbehaviorlab.com* and received an e-mail explaining the compensation and containing a link to the experiment. We predetermined that we wanted at least 100 participants, and once we observed that over 100 people had completed the experiment, we stopped data collection and analyzed the data.

This experiment used the same basic design as Bem's (2011) Experiment 8 with the following notable exceptions: It was conducted online (rather than in the lab) and used a different set of words in the same categories used by Bem.

Experiment 2

Participants ($n = 158$; 119 female, 39 male; median age = 39.5) were recruited from the same online panel and offered the same compensation as Experiment 1 (although none of the same individuals were in this sample). Again, participants received an e-mail that included the link to the experiment. We decided on a minimum sample of 150 for this experiment and stopped collecting data once we saw that we had passed that number.

This experiment used the same basic design as Bem's (2011) Experiment 8 with the following notable exceptions: It was conducted online (rather than in the lab) and used a different set of words taken from four different categories.

Experiment 3

Undergraduates ($n = 124$; 55 female, 69 male; median age = 19) at New York University participated in partial fulfillment of a course requirement. Each participant was scheduled to come into the lab, and upon arrival, was seated at a computer terminal and told to put on the available headphones. The experimenter opened the program, and participants went through the procedure at their own pace. We sought a sample of greater than 100 participants,

and because students are available in "batches" at NYU, we ended up with 124. This experiment used the same design and words as Bem's (2011) Experiment 8.

Experiment 4

Undergraduates (n = 109; 53 female, 55 male; 1 unknown; median age = 21) from Carnegie Mellon University and the University of California, Berkeley, participated for partial fulfillment of a course requirement. Scheduling and experimenter interaction were largely the same as in Experiment 3. We drew our sample from two universities because we wanted to make certain that we could reach a sample of at least 100 prior to the end of the semester, and neither participant pool could provide that many participants on its own. This experiment used the same words and design as Bem's (2011) Experiment 9.

Experiment 5

Undergraduates (n = 211; 116 female, 94 male, 1 unknown; median age = 20) from the University of Florida participated for extra course credit. Scheduling and experimenter interaction were largely the same as in Experiments 3 and 4. We sought a sample of at least 200. Because participants were scheduled in batches, we ended up with a number that was slightly higher. This experiment used the same words and design as Bem (2011) Experiment 9.

Experiment 6

Participants (n = 175; 122 female, 52 male, 1 unknown; median age = 36) were recruited from the same online panel as in Experiments 1 and 2. Again, participants received an e-mail that included the link to the experiment. Participants were assigned to one of two conditions. Some participants saw the same words and followed the same procedure as in Bem's (2011) Experiment 9 (Test-Before-Practice), whereas some received the same elements in the reverse order (Practice-Before-Test). This latter condition was included to establish that participants in an online sample are sufficiently attentive to benefit from practice (and thus, that any null results in Test-Before-Practice conditions could not be blamed on online participants failing to engage in practice). The Practice-Before-Test condition thus followed the sequence typically observed in memory experiments: participants answered the sensation-seeking items and watched a presentation of all 48 words. Then, 24 words were randomly selected by the computer (again, 6 from each of the 4 categories of 12 words), and participants watched a presentation of those 24 words and practiced the 24 words. Next, participants completed the free recall task of all 48 words, and finally, they reported whether or not they had paid attention during the experiment.

More people were intentionally assigned to the Test-Before-Practice condition than the Practice-Before-Test condition, and we left the program running until we observed that there were more than 100 people in the former condition: this led to 106 participants in the Test-Before-Practice condition and 69 in the Practice-Before-Test condition. The nonuniform random assignment was accomplished by having the computer program assign roughly one participant to the Practice-Before-Test condition for every two participants who completed the Test-Before-Practice condition.

This experiment, apart from the manipulation described above, used the same basic design as Bem's (2011) Experiment 9 but was conducted online (rather than in the lab).

Experiment 7

Participants (n = 2,469; demographic information not collected) were neither actively recruited nor compensated. After completing Experiment 1, the authors posted a short summary of that experiment on Social Science Research Network (SSRN), the online social science repository, and they included a link to an open study that could be completed by anyone with an Internet connection. A number of commentators on Bem (2011) also included hyperlinks to the short report. This, in turn, led to more people completing the open experiment. Data collection began on October 29th, 2010, and concluded on March 2nd, 2012 (when this article was written).

Data Coding Strategy

To assess whether or not we observed retroactive facilitation of recall, we first had to determine which words were recalled as a function of whether they were practiced. On the surface, this seems like a trivial task; however, there were occasionally spelling errors. For Experiments 1 and 2, we coded the recalled words in a two-stage process. First, all entered words that perfectly matched any of the 48 words from the set were coded as either coming from the practice set of words or coming from the control set of words (about 90% of all words fell into one of these two categories). This was done automatically by a computer program. Next, any listed words that did not match any of the 48 words from the set were manually checked, one at a time, to assess whether they were simply misspelled words (e.g., "spageti") or words that were not in the main set of words (e.g., "home"). In all cases, the determination of whether a word was a misspelling was entirely clear, and furthermore, in all cases, the coder was blind as to whether the words were drawn from the practice set or the control set.

For Experiments 3 through 7, we developed a fully computerized approach to coding the recalled words, thus removing any possible human bias in the scoring. Specifically, we used a computer program to generate exhaustive lists of common misspellings and typographical errors (e.g., "walruss" instead of "walrus"). If the recalled word matched any of the common misspellings, it was coded as a correctly recalled word.

Finally, for all experiments, any duplicate words were automatically identified and categorized as having come from the practice or control sets. Scores were adjusted accordingly (e.g., if the word "car" was in the control set and a participant responded with "car" twice, the second response was not counted as an additional recalled control word). The originally typed text, the lists of commonly misspelled words, and all of our data are freely available (http://www.consumerbehaviorlab.com/psi/CorrectingThePastData.xlsx).

Results

To test for the presence of precognition, Bem (2011) computed a weighted differential recall score (DR) for each participant using the formula

DR = (Recalled Practiced Words − Recalled Control Words)

× (Recalled Practice Words + Recalled Control Words).

In the article, for descriptive purposes, Bem (2011) frequently reported this number as DR%, which is the percentage that a participant's score deviated from random chance toward the highest or lowest scores possible (−576 to 576). We conducted the identical analysis on our data and also report DR% (see Table 1). In addition to using the weighted differential recall score, we also computed a simple unweighted recall score, which is the difference between recalled practice words and recalled control words (see Appendix B). For both of these measures, random chance would lead to a score of 0, and our analysis, like Bem's, was conducted using a one-sample *t* test.

Main Results

Table 1 presents the results of our seven experiments as well as the results of Bem's (2011) Experiments 8 and 9, for comparison. Bem found DR% = 2.27% in Experiment 8 and 4.21% in Experiment 9, effects that were significant at $p = .03$ and $p = .002$, one-tailed.

In contrast, only one of our seven experiments showed a significant effect suggesting precognition (using a one-tailed *p* value). Our seven experiments had an overall effect very close to zero.

In Experiment 1, DR% = −1.21%, $t(111) = -1.20$, $p = .88$.[1] Bayesian *t* tests suggest that this is "substantial" support for the null hypothesis of no precognition. Bayesian *t* tests (advocated by Wagenmakers et al., 2011) allow for hypothesis testing that considers the evidence for and against the null hypothesis, as well as the evidence for and against the alternative hypothesis. The analysis results in a Bayes Factor (BF) that denotes the weight of evidence provided by the data. Formally, the BF is computed as the probability of the data arising given H_0, over the probability of the data arising given H_1. When $BF > 1$, there is greater support for H_0, and when $0 < BF < 1$, there is greater support for H_1. For a more detailed review of Bayesian *t* tests, see Rouder, Speckman, Sun, Morey, and Iverson (2009).

In Experiment 2, DR% = 0.00%,[2] $t(157) = 0.00$, $p = .49$. Bayesian *t* tests suggest that this is "strong" support for the null hypothesis.

In Experiment 3, DR% = 1.17%, $t(123) = 1.28$, $p = .10$. Although DR% was indeed above zero, in the direction predicted by the ESP hypothesis, the test statistic did not reach conventional levels of significance, and Bayesian *t* tests suggest that this is nevertheless "substantial" support for the null hypothesis.

In Experiment 4, DR% = 1.59%, $t(108) = 1.77$, $p = .04$. The test statistic was significant in this one-tailed test, but Bayesian *t* tests suggest that this is "anecdotal" support for the null hypothesis.

In Experiment 5, DR% = −0.49%, $t(210) = -0.71$, $p = .76$. Bayesian *t* tests suggest that this is "strong" support for the null hypothesis.

In Experiment 6's Test-Before-Practice condition, DR% = −0.29%, $t(105) = -0.33$, $p = .63$. Bayesian *t* tests suggest that this is "strong" support for the null hypothesis.

In Experiment 7, which contained our largest sample of participants, DR% = −0.05%, $t(2468) = -0.23$, $p = .59$. Bayesian *t* tests suggest that this is "extreme" support for the null hypothesis.

In sum, in four of our experiments, participants recalled more control words than practice words (Experiments 1, 5, 6, and 7), and in three of our experiments, participants recalled more practice words than control words (Experiments 2, 3, and 4). One of these effects was statistically reliable using one-tailed *t* tests (see Table 1), but as noted, Bayesian *t* tests suggest that even the findings that were directionally consistent with precognition show substantial support for the null hypothesis of no precognition.

Practice-Before-Test, Experiment 6

In Experiment 6, we wanted to confirm that the basic underlying effect of practice-facilitated recall could be detected online. Accordingly, we assigned some participants to practice the words prior to the free recall test (a nonretroactive condition). In the Practice-Before-Test condition, the results were quite strong (DR% = 41.76%), $t(68) = 16.55$, $p < .001$. Not only was there a substantial mean difference between practiced and control words, but 68 of 69 participants recalled more practice words than control words (the remaining participant remembered the same number of each). Recall that in the same experiment, some participants received the precognition version (i.e., the retroactive condition). Despite coming from the same population and taking the experiment over the same medium, DR% did not differ reliably from zero in the retroactive condition, and in fact, participants remembered slightly more control words than practice words.

It is also worth noting that among the practice-before-test participants, people who recalled more words overall also showed a larger DR% ($r = .70$, $p < .001$). Even in this online environment, people who remembered more words (presumably reflecting more attention) also showed more benefits of practice, but only when the practicing preceded testing. When testing preceded practicing, this correlation was nonsignificant ($r = .01$, $p = .50$).

Sensation Seeking as a Correlate

In addition to the primary measure, Bem (2011) reported evidence suggesting that sensation seeking positively influenced precognitive ability. His evidence came in the form of a correlation between DR% and responses on the two-item sensation seeking scale. In Experiment 8, he reports a correlation of $r = .22$. In Experiment 9, the correlation drops to $r = -.10$, perhaps because "the same strong stimulus manipulation that produced the higher effect size also restricted the range of DR% scores sufficiently to squelch the predictive power of the individual difference measure" (Bem, 2011, p. 420). We did not observe a significant correlation across any of our experiments. Effect sizes ranged from $r = -.11$ in Experiment 4 to $r = .06$ in Experiment 6 (see Table 1). Sensation seeking did not predict (positively or negatively) precognitive performance in any of our experiments.

[1] To mirror the analysis conducted by Bem (2011), all *p* values for experimental data in this article are one-tailed in the positive direction, except where stated. Because we had no a priori predictions about moderators in the meta-analysis, all *p* values there are two-tailed.

[2] Throughout the article, we primarily report values to two significant digits. In some cases, this results in values of 0.00 and −0.00. In those cases, we include the sign to indicate that before rounding, the value is positive or negative.

Table 1
Experiment Results

Experiment	*N*	Mean DR%[a]	Statistic[b]	Bayesian *t* test	Correlation with sensation seeking
Bem (2011, Experiment 8)	100	2.27 (1.17)	$t(99) = 1.92, p = .03, d = 0.19$	BF = 2.11, Anecdotal (H_0)	$r = .22, p = .01$
Bem (2011, Experiment 9)	50	4.21 (1.41)	$t(49) = 2.96, p < .01, d = 0.42$	BF = 0.17, Substantial (H_1)	$r = -.10, p = .25$
Experiment 1	112	−1.21 (1.01)	$t(111) = -1.20, p = .88, d = 0.11$	BF = 6.58, Substantial (H_0)	$r = -.05, p = .71$
Experiment 2	158	0.00 (0.77)	$t(157) = 0.00, p = .49, d < 0.001$	BF = 15.85, Strong (H_0)	$r = -.06, p = .77$
Experiment 3	124	1.17 (0.92)	$t(123) = 1.28, p = .10, d = 0.11$	BF = 6.27, Substantial (H_0)	$r = -.03, p = .63$
Experiment 4	109	1.59 (0.90)	$t(108) = 1.77, p = .04, d = 0.17$	BF = 2.86, Anecdotal (H_0)	$r = -.11, p = .87$
Experiment 5	211	−0.49 (0.69)	$t(210) = -0.71, p = .76, d = 0.05$	BF = 14.23, Strong (H_0)	$r = -.01, p = .58$
Experiment 6 (test-before-practice)	106	−0.29 (0.88)	$t(105) = -0.33, p = .63, d = 0.03$	BF = 12.34, Strong (H_0)	$r = .06, p = .26$
Experiment 6 (practice-before-test)	69	41.76 (2.5)	$t(68) = 16.55, p < .001, d = 1.99$	BF < 0.01, Extreme (H_1)	$r = -.12, p = .85$
Experiment 7	2,469	−0.05 (0.22)	$t(2468) = -0.23, p = .59, d = -0.00$	BF = 60.66, Extreme (H_0)	$r = -.02, p = .81$
All psi data[c]	3,289	−0.04 (0.19)	$t(3287) = -0.20, p = .58, d < 0.01$	BF = 70.48, Extreme (H_0)	$r = -.08, p = .58$

Note. DR% = the percentage that a participant's score deviated from random chance toward the highest or lowest scores possible; BF = Bayes factor. [a] Values in parentheses are standard deviations. [b] All *p* values are one-tailed in the positive direction for DR% and for positive correlations for the correlational tests. [c] Includes data from all seven experiments except those in the practice-before-test condition in Experiment 6.

Meta-Analysis

In addition to conducting our own replications, another goal of this article was to examine all evidence for or against psi in the retroactive facilitation of recall paradigm. Accordingly, we conducted a meta-analysis of all known published and unpublished replication attempts of the two relevant experiments.

Retrieval of Studies

To locate all such attempts, we employed a number of different strategies. First, we searched for all articles that cite the original Bem (2011) article using Google Scholar, Web of Science, and ProQuest. We assumed that any attempts to replicate would cite Bem's article. Next, we posted a request for information regarding replication attempts on the following listservs: the *Society for Personality and Social Psychology,* the *Society of Experimental Social Psychology*, the *Society for the Psychological Study of Social Issues*, and the *Society for Judgment and Decision Making*. Additionally, we contacted the National Society of Paranormal Investigation and Research, the ParaPsychological Association, and the Society for Psychical Research, asking for any information about replication attempts by their constituents. Finally, because individual e-mail addresses were available, we directly contacted every member of the Rhine Research Center, the publishers of the *Journal of Parapsychology*. Some responders informed of us of individuals who may be conducting relevant replications, and we contacted all of those individuals. Every individual that we contacted who conducted a relevant study responded with either their data or with a description of their results.

Criteria for Selection of Studies

Our goal was to identify any direct replication attempts of either Experiment 8 or Experiment 9 from Bem (2011). To that end, we identified 12 replications and included 10 of them in our meta-analysis (see Table 2). We excluded two experiments reported by Snodgrass (2011) due to the limited sample size ($N = 1$ in Experiment 1, and $N = 9$ in Experiment 2). In addition, we included the original results obtained by Bem (2011) and the results from the seven experiments reported in this article. In total, this yielded data from 4,091 participants.

Calculation and Coding of Effect Sizes

Means and standard deviations were available for all replication attempts, and we calculated effect sizes (*d*) by dividing the DR% score by its standard deviation, with positive values indicating the presence of retroactive facilitation of recall and negative values indicating the presence of antiretroactive facilitation of recall. In addition to DR%, Bem (2011) reported a positive correlation between sensation seeking and DR% across all but the last of his nine experiments. Accordingly, we obtained these correlation estimates for the experiments in this meta-analysis either by extracting them from provided materials (e.g., published article or unpublished manuscripts) or by computing them ourselves using data provided by experimenters. We were unable to obtain this correlation for three replication attempts: Subbotsky (2012, Experiments 1 and 2) and Tressoldi, Masserdotti, and Marana (2012).

All effect sizes were coded on six dimensions: (a) whether the experiment attempted to replicate Bem's (2011) Experiment 8 or his Experiment 9, (b) whether it was administered online or in a lab, (c) whether it was conducted by Bem, (d) whether the software used to administer the experiment was the software originally used by Bem, (e) whether the results had already been published (we treat our results as unpublished), and (f) whether the experimenters conducting the replication expected to observe a psi effect.

The last criterion merits further explanation. Previous work has shown that experimental results can be influenced by experimenters' expectations (Rosenthal, 1966), and so we thought it appropriate to investigate whether psi effects might also be susceptible to such influence. Furthermore, it has been suggested that this type of expectancy might influence the operation of psi (D. J. Bem, personal communication, February 26, 2012). We were able to identify the experimenter expectation associated with each replication attempt by one of two means: (a) collecting publicly made statements by the experimenters (e.g., in their articles or on their

Table 2
Effect Sizes From All Replication Attempts

Experiment	*N*	Mean DR%[a]	*D*	*r*(DR%, sensation seeking)	Experiment type	Experiment administrator	Location of experiment	Software used	Publication status	Experimenter bias
Experiment 1	112	−1.21 (10.67)	−0.11	−.05	8	Not Bem	Online	Not Bem's	Unpublished	For
Experiment 2	158	0.00 (9.74)	<0.001	−.06	8	Not Bem	Online	Not Bem's	Unpublished	Against
Experiment 3	124	1.17 (10.25)	0.11	−.03	8	Not Bem	Lab	Not Bem's	Unpublished	Against
Experiment 4	109	1.59 (9.38)	0.17	−.11	9	Not Bem	Lab	Not Bem's	Unpublished	Against
Experiment 5	211	−0.49 (10.02)	−0.05	−.01	9	Not Bem	Lab	Not Bem's	Unpublished	Against
Experiment 6 (test-before-practice)	106	−0.29 (9.01)	−0.03	.06	9	Not Bem	Online	Not Bem's	Unpublished	Against
Experiment 7	2,469	−0.05 (10.99)	−0.00	−.02	8	Not Bem	Online	Not Bem's	Unpublished	Against
Bem, 2011, Exp 8	100	2.27 (11.75)	0.19	.22	8	Bem	Lab	Bem's	Published	For
Bem, 2011, Exp 9	50	4.21 (10.00)	0.42	−.10	9	Bem	Lab	Bem's	Published	For
Milyavsky, 2010[b]	58	−0.14 (13.82)	−0.01	−.12	9	Not Bem	Lab	Bem's	Unpublished	For
Pedersen et al., 2012	96	1.81 (9.61)	0.19	.00[e]	9	Not Bem	Lab	Bem's	Published	Against
Platzer, 2012[c]	98	1.29 (11.51)	0.11	−.09	9	Not Bem	Lab	Not Bem's	Unpublished	Against
Ritchie et al., 2012, Exp 1	50	0.19 (12.63)	0.01	.15	9	Not Bem	Lab	Bem's	Published	Against
Ritchie et al., 2012, Exp 2	50	−2.72 (12.23)	−0.22	−.19	9	Not Bem	Lab	Bem's	Published	Against
Ritchie et al., 2012, Exp 3	50	−0.58 (14.27)	−0.04	−.02	9	Not Bem	Lab	Bem's	Published	Against
Robinson, 2011	50	−1.60 (13.00)	−0.12	−.07	9	Not Bem	Lab	Bem's	Unpublished	For
Subbotsky, 2012, Exp 1	75	3.13 (11.08)	0.28		9	Not Bem	Lab	Bem's	Unpublished	For
Subbotsky, 2012, Exp 2	25	−3.06 (10.55)	−0.29		9	Not Bem	Lab	Bem's	Unpublished	For
Tressoldi et al., 2012[d]	100	2.25 (11.27)	0.20		9	Not Bem	Lab	Bem's	Unpublished	For

Note. DR% = the percentage that a participant's score deviated from random chance toward the highest or lowest scores possible; Exp = experiment. [a] Values in parentheses are standard deviations. [b] Experiment conducted using Hebrew words. [c] Experiment conducted using German words. [d] Experiment conducted using Italian words. [e] Sensation seeking was measured using a 40-item scale.

public blogs) or (b) contacting the experimenters and explicitly asking them what their expectation was. We coded the experiments that we conducted as follows. The lead investigator for Experiment 1 initially hypothesized that the experiment would yield positive results. Following the failure to replicate, the same investigator, falling in line with the remaining authors, subsequently updated his personal prior to that of obtaining a null result. It is worth noting that despite the fact that the authors of this article held priors about psi when conducting the experiments, the goal of our replication attempts was always to be as objective as possible. As far as we know, our expectations did not affect the programming of the experiments, data collection, or analyses. The expectation merely refers to the belief about psi that the experimenters held prior to conducting the experiments, not to a conscious agenda that was pursued.

Meta-Analysis of Effect Sizes

A summary of effect sizes is provided in Table 2 and Figure 1. To meta-analyze the effect sizes, we followed the procedure outlined by Hedges and Olkin (1985) and Lipsey and Wilson (2001). For DR%, we first adjusted the effect sizes to correct for biases associated with small samples (raw effect sizes are reported throughout the article). We then weighted the effect sizes by the inverse of the standard error of each point estimate to account for variations in sample size and then computed weighted average effect sizes for each level of our six effect size coding variables (see Table 3). For the correlation between DR% and sensation seeking, we first transformed all correlations using a Fisher's Zr transformation to compute correlation standard errors. Next, we weighted each Zr transformed correlation coefficient by $n - 3$ (Lipsey & Wilson, 2001) and computed weighted average correlations for each level of our six effect size coding variables.

DR%. The overall average effect size of .04 is considerably smaller than Bem's (2011) average effect size (.29) and is not statistically different from zero, 95% confidence interval (CI) [−.00, .09].

We next consider the effect of each of our coding variables separately. The average effect size of .02 for replications of Experiment 8 did not significantly differ from zero (95% CI [−.05, .08]), but the average effect size of .06 for replications of Experiment 9 did (95% CI [.01, .12]). The average effect size of .02 for replications not conducted by Bem (2011) did not significantly differ from zero, 95% CI [−.02, .07], but the average effect size of .29 for experiments conducted by Bem did, 95% CI [.13, .45]. The average effect size of −.02 for replications conducted online did not significantly differ from zero, 95% CI [−.09, .05], but the average effect size of .09 for experiments conducted in the lab did, 95% CI [.03, .14]. The average effect size of .01 for replications not using Bem's software did not significantly differ from zero, 95% CI [−.04, .07], but the average effect size of .09 for exper-

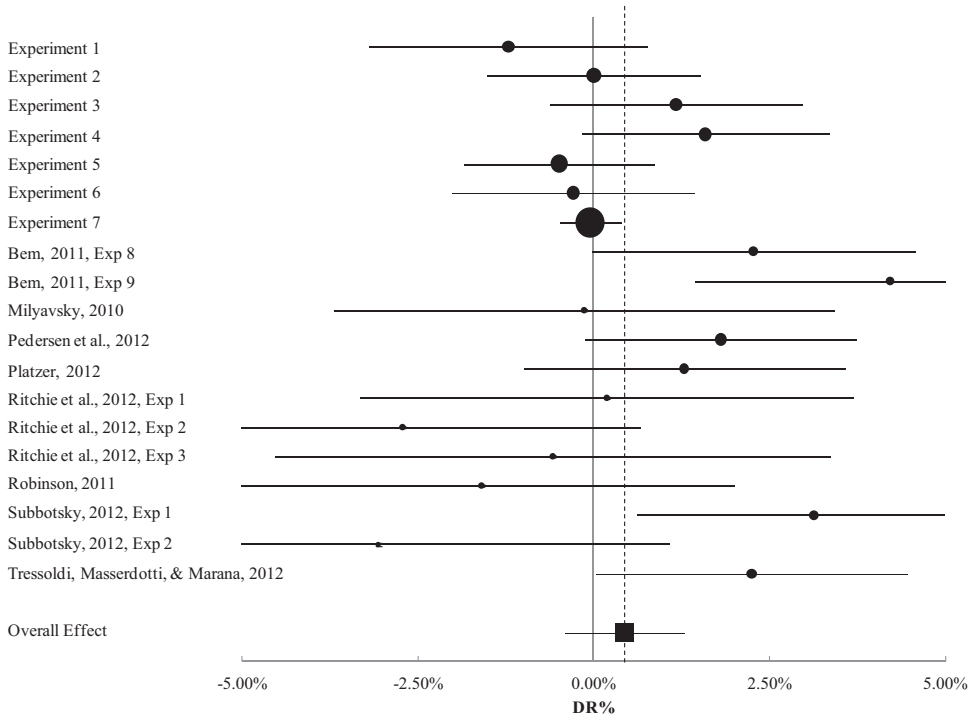

Figure 1. Forest plot of DR%. Size of circles represents the weight of the experiment in the meta-analysis. The vertical dotted line and square represent weighted average overall effect. Horizontal lines represent 95% confidence intervals. Exp = experiment; DR% = the percentage that a participant's score deviated from random chance toward the highest or lowest scores possible.

iments using Bem's software did, 95% CI [.02, .17]. The average effect size of .02 for unpublished replications did not significantly differ from zero, 95% CI [−.03, .07], but the average effect size of .12 for published replications did, 95% CI [.02, .22]. Finally, the average effect size of .03 for replications conducted by experimenters who did not expect to observe psi effects did not significantly differ from zero, 95% CI [−.03, .08], but the average effect size of .09 for replications conducted by experimenters who expected to observe psi effects did, 95% CI [.01, .17].

Despite these apparent differences, it is important to note that only one variable had a statistically significant influence on the size of the psi effect. That is, for only one potential moderator did the 95% CI around the point estimate of the differences in *d*s (between levels of the moderator) not include zero. This variable was whether or not the experiment was conducted by Bem (2011; difference in $d = 0.27$), 95% CI: [.10, .43]. The average effect size for experiments conducted by Bem is not only significantly different from zero, but it is also significantly higher than in replications conducted by anyone else. For the other moderators, this was not the case: The average effect size for replications of Experiment 8 did not significantly differ from replications of Experiment 9 (difference [diff] = −.05), 95% CI [−.14, .04], the average effect size for replications conducted online did not differ from replications conducted in a laboratory (diff = −.11), 95% CI [−.20, .00], the average effect size for experiments using Bem's software did not differ from experiments not using his software (diff = .08), 95% CI [−.01, .17], the average effect size for published replications did not differ from unpublished replications (diff = .10), 95% CI [−.01, .21], and the average effect size for replications conducted by researchers with positive expectations did not differ from replications conducted by researchers with negative expectations (diff = .07), 95% CI [−.03, .17].

It is also important to note that many of the moderators are highly correlated with each other and with whether Bem was the experimenter, and so many of the observed moderation effects likely do not represent unique effects. For example, in our sample, a study that is published also tends to be one that Bem conducted ($r = .46$), suggesting that the "Bem-as-experimenter" result may be driving the publication result. This is further confirmed by the fact that rerunning the meta-analysis with the 17 experiments ($N = 3{,}941$) not conducted by Bem results in every *d* becoming nonsignificantly different from zero. For example, when including Bem's (2011) original two experiments, positive experimenter expectancy yields a *d* of 0.09, 95% CI [0.01, 0.17], but excluding these two experiments yields a *d* of only 0.02, 95% CI [−0.08, 0.11]. The same can be said for experiments replicating Experiment 9, original $d = 0.06$, 95% CI [.01, .12]; new $d = 0.04$, 95% CI [−.02, .10], those done in a lab, original $d = 0.09$, 95% CI [.03, .14]; new $d = 0.06$, 95% CI [−.01, .12], those using Bem's software, original $d = 0.09$, 95% CI [.02, .17]; new $d = 0.04$, 95% CI [−.04, .12], and those that were published, original $d = 0.12$, 95% CI [.02, .22]; new $d = 0.00$, 95% CI [−.11, .11]. Given that this was the case for every dimension we examined, we conclude that the rather large effect sizes observed by Bem drove every potential moderator that our meta-analysis originally revealed.

Table 3
Effect Sizes by Category

Category	Total *N*	Effect sizes for DR%			Effect sizes for *r*(DR%, sensation seeking)		
		d	95% CI	Homogeneity[a]	*r*	95% CI	Homogeneity
Experiment type							
Experiment 8 ($n = 5$)	2,963	0.02	−0.05, 0.08	$Q(4) = 5.26, p = .26$	−.02	−0.05, 0.02	$Q(4) = 6.07, p = .19$
Experiment 9 ($n = 14$)	1,128	0.06	0.01, 0.12	$Q(13) = 32.64, p < .01$	−.08	−0.14, −0.01	$Q(10) = 4.64, p = .91$
Experiment administrator							
Bem ($n = 2$)	150	0.29	0.13, 0.45	$Q(1) = 1.95, p = .16$	.12	−0.05, 0.28	$Q(1) = 3.23, p = .06$
Not Bem ($n = 17$)	3,941	0.02	−0.02, 0.07	$Q(16) = 26.71, p = .05$	−.04	−0.07, 0.00	$Q(13) = 6.88, p = .91$
Location of experiment							
Online ($n = 4$)	2,845	−0.02	−0.09, 0.05	$Q(3) = 0.96, p = .81$	−.02	−0.06, 0.01	$Q(3) = 0.47, p = .93$
Lab ($n = 15$)	1,246	0.09	0.03, 0.14	$Q(14) = 32.60, p < .01$	−.05	−0.11, 0.02	$Q(11) = 12.70, p = .31$
Software used							
Bem's ($n = 11$)	704	0.09	0.02, 0.17	$Q(10) = 29.01, p < .01$	.01	−0.08, 0.10	$Q(7) = 9.15, p = .24$
Not Bem's ($n = 8$)	3,387	0.01	−0.04, 0.07	$Q(7) = 6.96, p = .43$	−.04	−0.07, 0.00	$Q(7) = 3.51, p = .83$
Publication status							
Published ($n = 5$)	396	0.12	0.02, 0.22	$Q(4) = 15.37, p < .01$	.04	−0.06, 0.14	$Q(4) = 7.63, p = .11$
Unpublished ($n = 14$)	3,791	0.02	−0.03, 0.07	$Q(13) = 20.44, p = .08$	−.04	−0.07, 0.00	$Q(10) = 3.94, p = .95$
Experimenter expectation							
For ($n = 7$)	520	0.09	0.01, 0.17	$Q(6) = 25.41, p < .01$	0.01	−0.10, 0.13	$Q(4) = 6.66, p = .16$
Against ($n = 12$)	3,571	0.03	−0.03, 0.08	$Q(11) = 11.60, p = .31$	−0.03	−0.07, 0.00	$Q(10) = 6.39, p = .78$
All							
Fixed effects ($n = 19$)[b]	4,091	0.04	−0.00, 0.09	$Q(18) = 38.97, p < .01$	−0.03	−0.06, 0.00	$Q(15) = 13.50, p = .56$
Random effects ($n = 19$)	4,091	0.05	−0.02, 0.12				

Note. CI = confidence interval; DR% = the percentage that a participant's score deviated from random chance toward the highest or lowest scores possible.
[a] The *p* values are one tailed. When $p < .05$, heterogeneity is assumed. [b] Because *r*(DR%, sensation seeking) for Subbotsky (2012) and Tressoldi et al. (2012) were not available, the *n* for that meta-analysis is only 16.

Sensation seeking. The average correlation between sensation seeking and DR% across all experiments was −.03, 95% CI [−0.06, 0.00], suggesting that there was no relationship between these two variables. Moreover, none of the variables we considered moderated this relationship, and we only observed one significant relationship in any of the subsets of these dimensions, a negative one, for experiments replicating Bem's (2011) Experiment 9. There seems to be insufficient evidence to conclude that sensation seeking correlates with psi.

Homogeneity. As can be seen in Table 3, the overall meta-analyses is heterogeneous, $Q(18) = 38.97, p < .01$, suggesting that a fixed effect meta-analytic model may be inappropriate. Accordingly, a random effects model was used that yielded nearly identical results. Specifically, the overall average effect size of 0.05 did not significantly differ from 0, 95% CI [−0.02, .12]. For simplicity, we do not report the average effect sizes using a random effects model for each level of moderator tested. However, the point estimates do not significantly vary as a function of the model used.

Because homogeneity was found for the overall sensation seeking analysis and for every level of moderator, a fixed effect model is sufficient, and so no random effects model was tested for sensation seeking.

Additional Analyses

Because Bem has made his data available (D. J. Bem, personal communication, November 1, 2010), we are able to perform additional analyses comparing his results with the results of our seven experiments. One way of comparing our results to Bem's is simply to test, via independent-sample *t* tests, whether the psi effect observed in our experiments was significantly *lower* than that observed in the original studies. When comparing our Experiments 1, 2, 3, and 7 against Bem's (2011) Experiment 8, we obtain the following results: $p = .03$, $p = .01$, $p = .47$, and $p = .04$, respectively. Comparing our Experiments 4, 5, and 6 against Bem's Experiment 9, we obtain the following results: $p = .11$, $p < .01$, and $p < .01$, respectively. With the exception of Experiments 3 and 4, all of our experiments produced a psi effect significantly lower than those reported by Bem.

Finally, because Experiment 7 differs greatly in sample size from all other experiments included in the meta-analysis, we reran the entire analysis excluding this experiment. As can be seen in Appendix C, with one exception, our conclusions do not greatly differ. When using a fixed effect model, the overall *d* of 0.06 does significantly differ from 0, 95% CI [.01, .11]. However, when controlling for heterogeneity with a random effects model, the corrected *d* of 0.05 does not significantly differ from 0, 95% CI [−.02, .13]. Accordingly, despite the rather large weight that Experiment 7 plays in the meta-analysis, excluding it does not meaningfully change the interpretation of our results. Moreover, the conclusions about the moderators are unchanged with Experiment 7 excluded. That is, the only moderator that yields significantly different results is whether the experiment was conducted by Bem or not. All other moderators do not yield statistically significant effects.

General Discussion

We conducted seven experiments testing for precognition and found no evidence supporting its existence. Participants were

asked to freely recall a set of words and then subsequently to practice them by retyping and categorizing them. Bem (2011) found (in two experiments with a total of 150 participants) that participants recalled more words from a set that they were then randomly assigned to practice. We did not find this. In our seven experiments (with 3,289 participants), participants were as likely to recall words that were subsequently practiced as words that were not subsequently practiced. Finally, in a meta-analysis including the results of all nine of these experiments (seven of ours and two of Bem's) and the results of 10 experiments conducted by other researchers, we observed an overall effect nonsignificantly different from zero ($d = 0.04$). This combination of results suggests that in the retroactive facilitation of recall paradigm, there is insufficient evidence to reject the null hypothesis. Additionally, we find no evidence to support a relationship between sensation seeking and psi ($r = -.03$).

Limitations

Despite our best efforts to conduct identical replications of Bem's (2011) Experiments 8 and 9, it is possible that the detection of psi requires certain methodological idiosyncrasies that we failed to incorporate into our experiments. For instance, after reading the replication packet provided by Bem (D. J. Bem, personal communication, November 1, 2010), we noticed that there were at least three differences between our experiments (which followed the procedure described in Bem's published article) and the full procedure actually employed by Bem. First, prior to the start of Bem's experiments, the experimenter was required to have a conversation with each participant in order to relax the participant. Second, prior to starting Bem's (2011) experiments, participants were asked two questions in addition to the sensation seeking scale (agreement with the statement, "I have lots of anxiety when I'm taking a test" and frequency of "have[ing] . . . practiced any form of meditation, self-hypnosis, relaxation exercises, or biofeedback"). Third, the set of words used by Bem were divided into common and uncommon words, something that we did not do in our Experiments 1, 2, and 7. Given the fragility of the observation of psi phenomena, it is possible that these methodological idiosyncrasies are necessary for reliable detection. Indeed, although we failed to replicate Bem's findings, we would be eager to know of a set of conditions that can reliably detect psi. That said, to the extent that Bem elected not to report these specific idiosyncrasies in his published article, we can only assume that he does not believe that they are necessary for the detection of psi.

Another limitation is in our choice of experiments to replicate and meta-analyze. Although, as mentioned, Bem's (2011) Experiments 8 and 9 make the most logical sense to replicate, our investigation into psi is limited to the (lack of) detection of retroactive facilitation of recall. We can reasonably claim a failure to observe this type of psi but can make no claims regarding precognition, retroactive priming, or retroactive habituation, the other three areas of psi investigated by Bem. For that, we call for more replication attempts by independent research teams.

Concerns About Online Samples

Of the seven experiments that we conducted, four were conducted online. It is not immediately clear why precognition would not be observed online (i.e., the theoretical development of the construct does not specify whether this should moderate the effect), but we thought that it was reasonable to give the online environment additional consideration. One possible concern might be that if people are taking the test at some remote location, their surroundings might be sufficiently distracting to make them less attentive.

In Appendix B, we report the outcome of two methods for excluding participants who were insufficiently attentive for Experiments 1, 2, and 6 and two additional measures for Experiment 7. One measure simply asked participants to self-report if they were not paying full attention. This measure appears to have some validity as that exclusion increased the measure of overall recall in all four online experiments. Nevertheless, it did not influence DR%. The second measure was behavioral: We recorded how long each participant spent on the task. We reasoned that participants who were working too quickly (or abandoning the experiment) were unlikely to have attended sufficiently to the task. We chose a relatively liberal cutoff and excluded any participant who was more than 1 standard deviation faster than the mean completion time. Again, this measure was validated in that the exclusion yielded a higher total recall score, but it had no noticeable influence on DR%. (For two experiments, it nonsignificantly increased DR%, and for two, it nonsignificantly decreased it.)

Because of the open nature of Experiment 7, additional precautions were taken to ensure data integrity. First, as described above, participants indicated whether or not they had previously taken this experiment or one like it in the past. Of the 2,469 participants, 250 indicated that they had. We analyze these data both with and without these participants and report the results in Appendix B. Second, because participants may have been interested in simply seeing what the experimental procedure was like, we identified participants who chose not to recall any words at all. Thirty-three participants did not recall any words, and again, to be conservative, we analyze the data both with and without them. Neither of these exclusion criteria had a discernible influence on the total number of words recalled or DR%.

Additionally, we analyzed whether DR% was influenced by the total number of words recalled, for both the online and the lab studies. The total number of words recalled can be seen as a reasonable proxy for how closely people attend to the stimuli. This measure was positively related to DR% in four studies and negatively related in the three others. It never approached significance in either online or lab studies.

Finally, one concern may be that participants actively sought to sabotage our experiments in the direction of observing a null result. Participants could have taken one of two strategies to undermine our investigation. First, they could have "recalled" either zero or all 48 words (something that could be accomplished by writing down the words as they appeared during the learning phase of the experiments). Either strategy would yield a DR% of 0. However, only 44 participants out of all 3,289 "recalled" zero words, and none "recalled" all 48, suggesting that this was not the case. Second, participants could have, a priori, decided to write down some subset of words as they were being displayed (say, the first 10) and only "recall" those words. Because practice and control words are randomly determined after the "recall" task, this strategy would, on average, also yield a DR% score of 0. Though we cannot empirically rule out this strategy, we can reason that it

would work best if the number of predetermined words to recall was even and not odd (i.e., an odd number of recalled words necessarily provides evidence either for or against psi). Following this strategy, the sinister participant could minimize the likelihood of contributing to the overall DR% score by recalling an even number of words. This, however, was not the case: There was no difference in the proportion of times the total number of words recalled was odd or even, $\chi^2(1, N = 3{,}289) = 0.11$, $p = .30$. Moreover, analyzing the results from only those participants who recalled an odd number of words yielded a DR% of −.30, $t(1394) = -0.99$, $p = .32$, suggesting that even when excluding participants who may have attempted to undermine our results in this way, we failed to observe psi. As such, we suspect that the nefariousness of our participants was minimal.

How Can These Results Be Reconciled With Bem (2011)?

Bem (2011) reported nine experiments ($n = 950$) suggesting that people can feel the future; we report seven experiments ($n = 3{,}289$) suggesting that people cannot. How is that possible? To start, it is certainly useful to point out that we are only looking at one basic procedure from the overall set of Bem experiments. Perhaps, it could be argued, precognition exists, but it cannot be detected in the retroactive facilitation of recall paradigm. Under that assumption, we might look at the original Bem article and suggest that Experiments 8 and 9 are simply Type I error—a false rejection of the null hypothesis. We do not have any empirical grounds for questioning the remaining seven experiments.

Still, even in Experiments 8 and 9, it is unclear how Bem (2011) could find significant support for a hypothesis that appears to be untrue. Elsewhere, critics of Bem have implicated his use of a one-tailed statistical test (Wagenmakers et al., 2011), testing multiple comparisons without correction (Wagenmakers et al., 2011), or perhaps simply a lurking file drawer with some less successful pilot experiments. All of these concerns fall under a larger category of researcher degrees of freedom, which raise the likelihood of falsely rejecting the null hypothesis (Simmons, Nelson, & Simonsohn, 2011). Some of these researcher degrees of freedom can be easily justified and have small and seemingly inconsequential effects. For example, Bem analyzes participant recall using an algorithm which weights the total number of correctly recalled words (i.e., DR%). He could have instead analyzed simple difference scores and found a similar, but not quite identical, result. Indeed, reanalyzing the data from Bem (2011), Experiment 9 still has a significant effect with this simpler scoring ($M = .96$); $t(49) = 2.46$, $p = .008$, one tailed, but Experiment 8 becomes nonsignificant ($M = .49$), $t(99) = 1.48$, $p = .071$, one tailed.

The scoring distinction is just a single example, but even for Bem's (2011) simple procedure, there are many others. For example, Bem's words are evenly split between common and uncommon words, a difference that was not analyzed (or reported) in the original article but may reflect an alternative way to consider the data: Perhaps psi only persists for uncommon words? He reports the results of his two-item sensation-seeking measure, but he does not analyze (or report collecting) additional measures of participant anxiety or experimenter-judged participant enthusiasm. Presumably, these were collected because there was a possibility that they may be influential as well, but when analyses revealed that they were not, they were dropped from the article. To be fair, because Bem reported two experiments on retroactive facilitation, his freedom is somewhat constrained. He cannot easily use DR% for one and a simple difference score for the other. On the other hand, he can certainly choose the one that works best for both studies and never report the other. Regardless, all of these decisions are defensible and possibly even recommended. Nevertheless, because their application is at the discretion of the researcher examining data after the completion of the experiment, they can make a true effect more difficult to discern. Researcher degrees of freedom do not make a finding false (e.g., the second law of thermodynamics is still true, even if a researcher tries multiple tests to detect it), but they do make it much harder to distinguish between truth and falseness in reported data.

Popper (1959/2002) defined a scientifically true effect as that "which can be regularly reproduced by anyone who carries out the appropriate experiment in the way prescribed" (pp. 23–24). Though decades have passed, that is still the operational definition of scientific truth. An effect is not an effect unless it is replicable, and a science is not a science unless it conducts (and values) attempted replications. No matter the outcome, it is indisputably admirable for Bem to encourage and facilitate the independent replication of his experiments. It is, by definition, what any scientist should do.

References

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

Carey, B. (2011a, January 6). Journal's paper on ESP expected to prompt outrage. *The New York Times,* p. A1.

Carey, B. (2011b, January 11). You might already know this. *The New York Times,* p. D1.

Dunne, B. J., & Jahn, R. G. (2003). Information and uncertainty in remote perception research. *Journal of Scientific Exploration, 17,* 207–241.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York, NY: Academic Press.

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science, 207,* 557–558. doi:10.1126/science.7352271

Lipsey, M. W., & Wilson, D. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Milton, J. (1997). Meta-analysis of free response ESP studies without altered states of consciousness. *Journal of Parapsychology, 61,* 279–319.

Milyavsky, M. (2010). [Failure to replicate Bem (2011) Experiment 9]. Unpublished raw data, Hebrew University of Jerusalem, Jerusalem, Israel.

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

Popper, K. (2002). *The logic of scientific discovery.* New York, NY: Routledge. (Original work published 1959)

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 retro-

active facilitation of memory recall. *Journal of the Society for Psychical Research, 75,* 142–147.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. East Norwalk, CT: Appleton-Century-Crofts.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi:10.3758/PBR.16.2.225

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: 1359–1366, 2011. doi:10.1177/0956797611417632

Snodgrass, S. (2011). Examining retroactive facilitation of recall: An adapted replication of Bem (2011, Study 9) and Galak and Nelson (2010). *Social Science Research Network.* Retrieved from http://ssrn.com/abstract=1935942

Storm, L., Tressoldi, P. E., & Di Risio, L. D. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin, 136,* 471–485. doi:10.1037/a0019457

Subbotsky, E. (2012). *Sensing the future: The non-standard observer effect on an ESP task.* Lancaster, England: Lancaster University.

Tressoldi, P. E. (2011). Extraordinary claims require extraordinary evidence: The case of non-local perception, a classical and Bayesian review of evidences. *Frontiers in Psychology, 2,* 1–5. doi:10.3389/fpsyg.2011.00117

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.* Padua, Italy: Universita di Padova.

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

Appendix A

Lists of Words Used in Experiments

Table A1
List of Words Used by Category in Experiment 1 and 7

Food	Animals	Occupations	Clothes
apple	alligator	accountant	coat
bagel	cat	athlete	dress
bread	cow	bartender	hat
hamburger	dog	doctor	jeans
lasagna	dolphin	engineer	pants
omelet	frog	fireman	shirt
orange	goat	fisherman	shoes
pizza	horse	janitor	shorts
salad	lion	musician	skirt
sandwich	monkey	plumber	socks
spaghetti	pig	policeman	suit
steak	rabbit	teacher	underwear

Table A2
List of Words Used by Category in Experiment 2

Kitchen items	Electronics	Body parts	Sports
blender	calculator	chest	baseball
bowl	camera	ear	basketball
dishwasher	cellphone	eye	bat
fork	clock	finger	bicycle
knife	computer	foot	football
microwave	headphones	hand	goal
oven	printer	head	helmet
refrigerator	projector	knee	hoop
spatula	radio	mouth	puck
spoon	speakers	nose	skate
stove	stereo	shoulder	ski
toaster	television	toe	snowboard

(*Appendices continue*)

Table A3
List of Words Used by Category in Experiments 3–6 (From Bem 2011)

Food	Animals	Occupations	Clothes
apple	bird	bricklayer	bikini
bagel	cat	carpenter	coat
bread	chipmunk	comedian	dress
caviar	cow	doctor	hat
hamburger	dog	engineer	jockstrap
oatmeal	gorilla	lawyer	pantyhose
onion	horse	mortician	parka
potato	kangaroo	nun	shirt
soup	ostrich	nurse	shoes
tofu	skunk	rabbi	shorts
turnip	snake	scientist	suspenders
yogurt	walrus	teacher	tuxedo

(*Appendices continue*)

Appendix B

Table B1
Full Results

Experiment	*N*	P	C	Mean DR%	Weighted differential recall		Simple differential recall		
					Statistic	Bayesian *t* test	Mean	Statistic	Bayesian *t* test
Bem (2011, Experiment 8)	**100**	**9.42 (.35)**	**8.93 (.28)**	**2.27% (1.17)**	***t*(99) = 1.92, *p* = .03**	**BF = 2.11, Anecdotal (H_0)**	**.49 (.33)**	***t*(99) = 1.48, *p* = .14**	**BF = 4.32, Substantial (H_0)**
Bem (2011, Experiment 9)	**50**	**11.32 (.45)**	**10.36 (.36)**	**4.21% (1.41)**	***t*(49) = 2.96, *p* < .01**	**BF = 0.17, Substantial (H_1)**	**.96 (.39)**	***t*(49) = 2.46, *p* = .02**	**BF = 0.55, Substantial (H_1)**
Experiment 1: Full sample	**112**	**8.04 (.36)**	**8.36 (.38)**	**−1.21% (1.01)**	***t*(111) = −1.20, *p* = .88**	**BF = 6.58, Substantial (H_0)**	**−.31 (.30)**	***t*(111) = −1.06, *p* = .86**	**BF = 7.69, Substantial (H_0)**
Removing self−identified inattentive participants	104	8.25 (.38)	8.52 (.40)	−1.34% (1.11)	*t*(103) = −1.10, *p* = .86	BF = 7.11, Substantial (H_0)	−.27 (.31)	*t*(103) = −0.86, *p* = .81	BF = 8.952, Substantial (H_0)
Removing participants who were too fast on the recall portion (<1 *SD*)	103	8.54 (.34)	8.91 (.37)	−1.37% (1.09)	*t*(102) = −1.25, *p* = .87	BF = 5.96, Substantial (H_0)	−.37 (.32)	*t*(102) = −1.16, *p* = .88	BF = 6.62, Substantial (H_0)
Removing people from either of those two categories.	95	8.81 (.36)	9.14 (.38)	−1.35% (1.18)	*t*(94) = −1.15, *p* = .87	BF = 6.44, Substantial (H_0)	−.33 (.34)	*t*(94) = −0.96, *p* = .83	BF = 7.83, Substantial (H_0)
Experiment 2: Full sample	**158**	**8.24 (.34)**	**8.20 (.35)**	**0.00% (0.77)**	***t*(157) = 0.00, *p* = .49**	**BF = 15.85, Strong (H_0)**	**.06 (.24)**	***t*(157) = 0.24, *p* = .40**	**BF = 15.41, Strong (H_0)**
Removing self−identified inattentive participants	139	8.30 (.35)	8.38 (.36)	−0.63% (0.89)	*t*(138) = −0.42, *p* = .33	BF = 13.63, Strong (H_0)	−.08 (.25)	*t*(138) = −0.32, *p* = .38	BF = 14.14, Strong (H_0)
Removing participants who were too fast on the recall portion (<1 *SD*)	137	9.15 (.32)	9.07 (.34)	0.06% (0.89)	*t*(136) = 0.07, *p* = .47	BF = 14.74, Strong (H_0)	.09 (.27)	*t*(136) = 0.33, *p* = .37	BF = 14.00, Strong (H_0)
Removing participants from either of those two categories	124	9.01 (.34)	9.04 (.349)	−.32% (0.91)	*t*(123) = −0.35, *p* = .36	BF = 13.24, Strong (H_0)	−.03 (.27)	*t*(123) = −0.12, *p* = .46	BF = 13.97, Strong (H_0)
Experiment 3: Full sample	**124**	**8.51 (.32)**	**8.18 (.30)**	**1.17% (0.92)**	***t*(123) = 1.28, *p* = .10**	**BF = 6.27, Substantial (H_0)**	**.33 (.29)**	*t*(123) = 1.16, *p* = .15	**BF = 7.24, Substantial (H_0)**
Removing participants who were too fast on the recall portion (<1 *SD*)	111	9.16 (.30)	8.72 (.29)	1.43% (1.02)	*t*(110) = 1.40, *p* = .08	BF = 5.08, Substantial (H_0)	.44 (.31)	*t*(110) = 1.43, *p* = .08	BF = 4.88, Substantial (H_0)
Experiment 4: Full sample	**109**	**8.36 (.34)**	**7.83 (.33)**	**1.59% (0.90)**	***t*(108) = 1.77, *p* = .04**	**BF = 2.86, Anecdotal (H_0)**	**.53 (.29)**	***t*(108) = 1.82, *p* = .04**	**BF = 2.62, Anecdotal (H_0)**
Removing participants who were too fast on the recall portion (<1 *SD*)	98	8.82 (.34)	8.36 (.32)	1.29% (0.99)	*t*(97) = 1.60, *p* = .06	BF = 3.58, Substantial (H_0)	.46 (.31)	*t*(97) = 1.48, *p* = .07	BF = 4.29, Substantial (H_0)
Experiment 5: Full sample	**211**	**8.62 (.22)**	**8.64 (.25)**	**−0.49% (0.69)**	***t*(210) = −0.71, *p* = .76**	**BF = 14.23, Strong (H_0)**	**−.03 (.22)**	***t*(210) = −0.15, *p* = .56**	**BF = 18.08, Strong (H_0)**
Removing participants who were too fast on the recall portion (<1 *SD*)	186	9.14 (.22)	9.22 (.24)	−0.57% (0.77)	*t*(185) = −0.75, *p* = .77	BF = 12.99, Strong (H_0)	−.08 (.23)	*t*(185) = −0.32, *p* = .63	BF = 16.33, Strong (H_0)
Experiment 6: Full sample (test−before−practice)	**106**	**7.75 (.46)**	**8.08 (.42)**	**−0.29% (0.88)**	***t*(105) = −0.33, *p* = .63**	**BF = 12.34, Strong (H_0)**	**−.32 (.27)**	***t*(105) = −1.21, *p* = .89**	**BF = 6.33, Substantial (H_0)**
Removing self−identified inattentive participants	98	8.05 (.47)	8.37 (.43)	−0.30% (0.93)	*t*(97) = −0.33, *p* = .63	BF = 11.87, Strong (H_0)	−.32 (.28)	*t*(97) = −1.13, *p* = .88	BF = 6.68, Substantial (H_0)
Removing people who were too fast on the recall portion (<1 *SD*)	97	8.36 (.45)	8.63 (.41)	−0.24% (0.96)	*t*(96) = −0.26, *p* = .60	BF = 12.06, Strong (H_0)	−.27 (.29)	*t*(96) = −0.92, *p* = .82	BF = 8.21, Substantial (H_0)
Removing people from either of those two categories.	90	8.64 (.45)	8.92 (.41)	−0.26% (1.01)	*t*(89) = −0.25, *p* = .60	BF = 11.65, Strong (H_0)	−.28 (.30)	*t*(89) = −0.92, *p* = .82	BF = 7.92, Substantial (H_0)
Experiment 6: Full sample (practice−before−test)[a]	**69**	**15.61 (.61)**	**3.70 (.46)**	**47.76% (2.52)**	***t*(68) = 16.55, *p* < .001**	**BF < 0.01, Extreme (H_1)**	**11.58 (0.59)**	***t*(66) = 19.54, *p* < .001**	**BF < 0.01, Extreme (H_1)**
Removing self−identified inattentive people	62	15.73 (.66)	3.87 (.50)	42.23% (2.73)	*t*(61) = 15.49, *p* < .001	BF < 0.01, Extreme (H_1)	11.85 (0.55)	*t*(61) = 21.45, *p* < .001	BF < 0.01, Extreme (H_1)
Experiment 7: Full sample	**2,469**	**8.93 (.08)**	**8.96 (.08)**	**−.05% (.22)**	***t*(2,468) = −0.23, *p* = .59**	**BF = 60.66, Extreme (H_0)**	**−.03 (.06)**	***t*(2,468) = −0.52, *p* = .70**	**BF = 54.42, Extreme (H_0)**
Removing self−identified inattentive participants	2,145	9.12 (.09)	9.15 (.09)	−.04% (.24)	*t*(2,144) = −0.16, *p* = .57	BF = 57.33, Extreme (H_0)	−.04 (.07)	*t*(2,145) = −0.52, *p* = .70	BF = 50.73, Extreme (H_0)
Removing self−identified repeat participants	2,219	8.86 (.08)	8.91 (.08)	−.13% (.23)	*t*(2,218) = −0.58, *p* = .72	BF = 49.93, Extreme (H_0)	−.05 (.07)	*t*(2,218) = −0.78, *p* = .79	BF = 43.58, Extreme (H_0)
Removing participants with zero recalled words	2,436	9.05 (.08)	9.08 (.08)	−.05% (.22)	*t*(2,435) = −0.23, *p* = .59	BF = 60.26, Extreme (H_0)	−.03 (.06)	*t*(2,435) = −0.52, *p* = .70	BF = 54.06, Extreme (H_0)
Removing participants who were too fast on the recall portion (<1 *SD*)	2,420	9.10 (.08)	9.13 (.08)	−.05% (.23)	*t*(2,419) = −0.21, *p* = .59	BF = 60.33, Extreme (H_0)	−.03 (.06)	*t*(2,419) = −0.40, *p* = .69	BF = 56.94, Extreme (H_0)
Removing participants from any of those four categories	1,913	9.14 (.09)	9.20 (.09)	−.14% (.26)	*t*(1,911) = −0.54, *p* = .71	BF = 43.52, Extreme (H_0)	−.06 (.07)	*t*(1,911) = −0.82, *p* = .80	BF = 35.98, Extreme (H_0)

(*Appendices continue*)

Table B1 (*continued*)

Experiment	*N*	P	C	Mean DR%	Percentage of participants differentially recalling practice and control words			Correlation with sensation seeking	Correlation between DR% and total recall[b]
					P > C	P = C	P < C		
Bem (2011, Experiment 8)	**100**	**9.42 (.35)**	**8.93 (.28)**	**2.27% (1.17)**	**47.0% (47 of 100)**	**15.0% (15 of 100)**	**38.0% (38 of 100)**	***r* = .22, *p* = .014**	***r* = .26, *p* = .01**[b]
Bem (2011, Experiment 9)	**50**	**11.32 (.45)**	**10.36 (.36)**	**4.21% (1.41)**	**62.0% (31 of 50)**	**8.0% (4 of 50)**	**30.0% (15 of 50)**	***r* = −.10, *p* = .25**	***r* = −.33, *p* = .02**
Experiment 1: Full sample	**112**	**8.04 (.36)**	**8.36 (.38)**	**−1.21% (1.01)**	**38.4% (43 of 112)**	**13.4% (15 of 112)**	**48.2% (54 of 112)**	***r* = −.05, *p* = .71**	***r* = −.15, *p* = .11**
Removing self−identified inattentive participants	104	8.25 (.38)	8.52 (.40)	−1.34% (1.11)	39.4% (41 of 104)	13.5% (14 of 104)	47.1% (49 of 104)	*r* = −.06, *p* = .71	*r* = −.16, *p* = .11
Removing participants who were too fast on the recall portion (<1 *SD*)	103	8.54 (.34)	8.91 (.37)	−1.37% (1.09)	37.9% (39 of 103)	12.6% (13 of 103)	49.5% (51 of 103)	*r* = −.06, *p* = .71	*r* = −.15, *p* = .12
Removing people from either of those two categories.	95	8.81 (.36)	9.14 (.38)	−1.35% (1.18)	38.9% (37 of 95)	12.6% (12 of 95)	48.4% (46 of 95)	*r* = −.57, *p* = .71	*r* = −.16, *p* = .12
Experiment 2: Full sample	**158**	**8.24 (.34)**	**8.20 (.35)**	**0.00% (0.77)**	**41.1% (65 of 158)**	**10.7% (17 of 158)**	**48.1% (76 of 158)**	***r* = −.06, *p* = .77**	***r* = −.06, *p* = .43**
Removing self−identified inattentive participants	139	8.30 (.35)	8.38 (.36)	−0.63% (0.89)	38.8% (54 of 139)	10.1% (14 of 139)	51.1% (71 of 139)	*r* = −.05, *p* = .72	*r* = −.07, *p* = .43
Removing participants who were too fast on the recall portion (<1 *SD*)	137	9.15 (.32)	9.07 (.34)	.06% (0.89)	41.6% (57 of 137)	10.9% (15 of 137)	47.5% (65 of 137)	*r* = −.06, *p* = .74	*r* = −.09, *p* = .31
Removing participants from either of those two categories	124	9.01 (.34)	9.04 (.349)	−.32% (0.91)	39.5% (49 of 124)	10.5% (13 of 124)	50.0% (62 of 124)	*r* = −.05, *p* = .71	*r* = −.08, *p* = .36
Experiment 3: Full sample	**124**	**8.51 (.32)**	**8.18 (.30)**	**1.17% (0.92)**	**44.4% (55 of 124)**	**16.1% (20 of 124)**	**39.5% (49 of 124)**	***r* = −.03, *p* = .63**	***r* = .10, *p* = .29**
Removing participants who were too fast on the recall portion (<1 *SD*)	111	9.16 (.30)	8.72 (.29)	1.43% (1.02)	46.8% (52 of 111)	11.7% (13 of 111)	41.4% (46 of 111)	*r* = −.03, *p* = .61	*r* = .07, *p* = .46
Experiment 4: Full sample	**109**	**8.36 (.34)**	**7.83 (.33)**	**1.59% (0.90)**	**48.6% (53 of 109)**	**13.8% (15 of 109)**	**37.6% (41 of 109)**	***r* = −.11, *p* = .87**	***r* = .15, *p* = .12**
Removing participants who were too fast on the recall portion (<1 *SD*)	98	8.82 (.34)	8.36 (.32)	1.29% (0.99)	48.0% (47 of 98)	13.3% (13 of 98)	38.7% (38 of 98)	*r* = −.11, *p* = .87	*r* = .17, *p* = .09
Experiment 5: Full sample	**211**	**8.62 (.22)**	**8.64 (.25)**	**−0.49% (0.69)**	**43.6% (92 of 211)**	**11.8% (25 of 211)**	**44.5% (94 of 211)**	***r* = −.01, *p* = .58**	***r* = −.09, *p* = .17**
Removing participants who were too fast on the recall portion (<1 *SD*)	186	9.14 (.22)	9.22 (.24)	−0.57% (0.77)	42.5% (79 of 186)	11.8% (22 of 186)	45.7% (85 of 186)	*r* = −.01, *p* = .54	*r* = −.10, *p* = .18
Experiment 6: Full sample (test−before−practice)	**106**	**7.75 (.46)**	**8.08 (.42)**	**−0.29% (0.88)**	**34.0% (36 of 106)**	**22.6% (24 of 106)**	**43.4% (46 of 106)**	***r* = .06, *p* = .26**	***r* = .12, *p* = .22**
Removing self−identified inattentive participants	98	8.05 (.47)	8.37 (.43)	−0.30% (0.93)	32.7% (32 of 98)	23.5% (23 of 98)	43.8% (43 of 98)	*r* = .05, *p* = .30	*r* = .13, *p* = .22
Removing people who were too fast on the recall portion (<1 *SD*)	97	8.36 (.45)	8.63 (.41)	−0.24% (0.96)	37.1% (36 of 97)	19.6% (19 of 97)	43.3% (42 of 97)	*r* = .07, *p* = .25	*r* = .13, *p* = .21
Removing people from either of those two categories	90	8.64 (.45)	8.92 (.41)	−0.26% (1.01)	35.6% (32 of 90)	20.0% (18 of 90)	44.4% (40 of 90)	*r* = .06, *p* = .30	*r* = .14, *p* = .19
Experiment 6: Full sample (practice−before−test)[a]	**69**	**15.61 (.61)**	**3.70 (.46)**	**41.76% (2.52)**	**98.6% (68 of 69)**	**1.4% (1 of 69)**	**0%**	***r* = −.12, *p* = .85**	***r* = .70, *p* < .001**
Removing self−identified inattentive people	62	15.73 (.66)	3.87 (.50)	42.23% (2.73)	98.4% (61 of 62)	1.6% (1 of 62)	0%	*r* = −.15, *p* = .88	*r* = .69, *p* < .001
Experiment 7: Full sample	**2,469**	**8.93 (.08)**	**8.96 (.08)**	**−.05% (.22)**	**42.5% (1,049 of 2,469)**	**15.6% (385 of 2,469)**	**41.9% (1,035 of 2,469)**	***r* = −.02, *p* = .81**	***r* = .02, *p* = .33**
Removing self−identified inattentive participants	2,145	9.12 (.09)	9.15 (.09)	−.04% (.24)	42.8% (918 of 2,145)	14.8% (318 of 2,145)	42.4% (909 of 2,145)	*r* = −.02, *p* = .78	*r* = .03, *p* = .14
Removing self−identified repeat participants	2,219	8.86 (.08)	8.91 (.08)	−.13% (.23)	42.5% (943 of 2,219)	14.8% (329 of 2,219)	42.7% (947 of 2,219)	*r* = −.01, *p* = .65	*r* = .01, *p* = .63
Removing participants with zero recalled words	2,436	9.05 (.08)	9.08 (.08)	−.05% (.22)	43.1% (1,049 of 2,436)	14.4% (352 of 2,436)	42.5% (1,035 of 2,436)	*r* = −.02, *p* = .80	*r* = .02, *p* = .30
Removing participants who were too fast on the recall portion (<1 *SD*)	2,420	9.10 (.08)	9.13 (.08)	−.05% (.23)	43.1% (1,042 of 2,420)	14.5% (351 of 2,420)	42.4% (1,027 of 2,420)	*r* = −.02, *p* = .81	*r* = .02, *p* = .30
Removing participants from any of those four categories	1,913	9.14 (.09)	9.20 (.09)	−.14% (.26)	43.0% (823 of 1,913)	13.6% (261 of 1,913)	43.3% (829 of 1,913)	*r* = −.01, *p* = .61	*r* = .02, *p* = .35

Note. The four means presented in this table (P, C, DR%, and simple differential recall) are each presented with the standard error reported in parentheses. Bold indicates analyses on complete samples from the respective experiments. DR% = the percentage that a participant's score deviated from random chance toward the highest or lowest scores possible; P = the number of practice words correctly recalled (out of 24 possible); C = the number of control words correctly recalled (out of 24 possible); BF = Bayes factor.

[a] No participants were faster than a standard deviation from the mean in Experiment 6 (practice-before-test). [b] Because total number of words recalled were not provided by Bem, totals for experiments conducted by Bem are calculated as Practice Words Recalled + Control Words Recalled and may exclude words listed that were not part of the practice or control word sets. Additionally, because there is no a priori hypothesis regarding the direction of this correlation, *p* values are two-tailed.

(*Appendices Continue*)

Appendix C

Table C1
Effect Sizes by Category Excluding Experiment 7

Category	Total *N*	Effect sizes for DR%			Effect sizes for *r*(DR%, sensation seeking)		
		d	95% CI	Homogeneity[a]	*r*	95% CI	Homogeneity
Experiment type							
Experiment 8 ($n = 4$)	494	0.04	−0.05, 0.14	$Q(3) = 4.82, p = .19$	.01	−0.08, 0.10	$Q(3) = 5.79, p = .12$
Experiment 9 ($n = 14$)	1,128	0.06	0.01, 0.12	$Q(13) = 32.64, p < .01$	−.08	−0.14, −0.01	$Q(9) = 4.64, p = .86$
Experiment administrator							
Bem ($n = 2$)	150	0.29	0.13, 0.45	$Q(1) = 1.95, p = .16$	.12	−0.05, 0.28	$Q(1) = 3.32, p = .07$
Not Bem ($n = 16$)	1,472	0.03	−0.02, 0.08	$Q(15) = 26.27, p = .04$	−.07	−0.12, −0.01	$Q(12) = 4.92, p = .96$
Location of experiment							
Online ($n = 3$)	376	−0.04	−0.15, 0.06	$Q(2) = 0.69, p = .71$	−.06	−0.16, 0.04	$Q(2) = 0.00, p = .99$
Lab ($n = 15$)	1,246	0.09	0.03, 0.14	$Q(14) = 32.60, p < .01$	−.05	−0.11, 0.02	$Q(11) = 12.7, p = .31$
Software used							
Bem's ($n = 11$)	704	0.09	0.02, 0.17	$Q(10) = 29.01, p < .01$	.01	−0.08, 0.10	$Q(7) = 9.15, p = .24$
Not Bem's ($n = 7$)	918	0.02	−0.05, 0.09	$Q(6) = 6.74, p = .35$	−.08	−0.15, −0.01	$Q(6) = 1.11, p = .98$
Publication status							
Published ($n = 5$)	396	0.12	0.02, 0.22	$Q(4) = 15.37, p < .01$	.04	−0.06, 0.14	$Q(4) = 7.63, p = .11$
Unpublished ($n = 13$)	1,322	0.03	−0.02, 0.09	$Q(12) = 19.98, p = .07$	−.08	−0.14, −0.02	$Q(9) = 1.2, p = .99$
Experimenter expectation							
For ($n = 7$)	570	0.09	0.01, 0.17	$Q(7) = 25.41, p < .01$	.00	−0.10, 0.11	$Q(4) = 6.66, p = .16$
Against ($n = 11$)	1,052	0.04	−0.03, 0.10	$Q(9) = 11.08, p = .27$	−.07	−0.13, −0.01	$Q(9) = 4.74, p = .86$
All							
Fixed effects ($n = 18$)[b]	1,622	0.06	0.01, 0.11	$Q(17) = 37.63, p < .01$	−.05	−0.10, 0.00	$Q(14) = 12.74, p = .55$
Random effects ($n = 18$)	1,622	0.05	−0.02, 0.13				

Note. CI = confidence interval; DR% = the percentage that a participant's score deviated from random chance toward the highest or lowest scores possible.
[a] *p*-values are one-tailed. When $p < .05$, heterogeneity is assumed. [b] Because *r*(DR%, sensation seeking) for Subbotsky (2012) and Tressoldi et al. (2012) were not available, the *n* for that meta-analysis is only 16.

Received February 8, 2012
Revision received June 19, 2012
Accepted June 19, 2012 ■