A replication of the procedures from Bem (2010, Study 8) and a failure to replicate the same results.

Jeff Galak

Carnegie Mellon University

Leif D. Nelson

University of California, Berkeley

First Posted Online: October 29, 2010

Most Recently Updated: October 29, 2010

Address correspondence to either author at jgalak@cmu.edu or leif_nelson@haas.berkeley.edu.

Abstract

We replicated the procedure of Experiment 8 from Bem (2010), which had originally demonstrated retroactive facilitation of recall. We failed to replicate the result. The paper includes a description of our procedure and analysis as well as a brief discussion for some reasons why we obtained a different result than in the original paper.

A replication of the procedures from Bem (2010, Study 8) and a failure to replicate the same results.

Recently, Bem (2010) published an extremely thought-provoking article demonstrating the existence of precognition, "the conscious cognitive awareness… of a future event that could not otherwise be anticipated through any known inferential process." Through nine meticulously constructed experiments, using a range of tasks, Bem finds consistent support for the idea that people have precognitive abilities. As Bem suggests, the purpose of the paper was not exclusively to simply report evidence relevant to precognition, but also to develop procedures "that can be replicated by independent investigators (p. 3)". We sought out to replicate one of those procedures.

For our experiment we chose to do a replication of Experiment 8, the retroactive facilitation of recall. Below we detail the exact procedure, but in a rough sketch, people were shown a list of words and then asked to freely recall as many as possible. Participants were then randomly assigned to practice half of the list of words. Evidence of precognition would be observed if people freely recalled more of the words that they subsequently practiced than of the words that they subsequently did not practice.

It is worth giving a brief consideration to why we chose that particular paradigm. Seven of the nine experiments hinge on an affective response; arousal to erotic images, a preference to avoid a negative image, etc. There is nothing wrong with any of those procedures, but they require some judgment calls from the experimenter in selecting stimuli. As Bem reports, for example, finding stimuli that lead to sensitization or habituation can be tricky. It requires pretesting, of course, but it also leaves open a pretty easy explanation for a null finding: "maybe the stimuli weren't chosen correctly to be sensitive to the presence or absence of the effect." The comparatively simple retroactive memory experiments (8-9) seem to offer an easier set-up: choose 12 familiar words from four categories and let randomization take care of the rest. Even this procedure has its problems however. Most notably, humans are helpful in scoring the output of the participants. This raises two issues. First, it requires a little bit more labor from the experimenter, and therefore makes it a little more difficult to run larger samples or run multiple tests. Second, it leaves room for bias. As far as we can tell, this is trivial or non-existent. A participant who recalls the word *Apple* but types it as *Aple*, has clearly recalled the word. Furthermore, as we describe below, all coding is done blind to whether the words were from the practice or control sets, and is therefore unbiased. Nevertheless, given that the history of Psi research is marked by subtle influences of experimenter bias, it would be preferable to have a procedure which removed even that small human element from the analysis.

Method

Participants from an online participant pool (n = 112; Median Age = 38, 88 Females, 73% White, 7% Chinese, 5% Black or African American, 5% Hispanic, 10% Other) were recruited to participate in an experiment on extrasensory perception (ESP). Participants were compensated by earning entry into a lottery for $100, a standard incentive offered to participants in this pool[1].

Participants first read and agreed to a consent form mentioning again that the study was investigating ESP and then read a brief introductory statement almost identical to the one used by Bem (2010). “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), they clicked continue and proceeded to the next screen.

On the two subsequent screens participants answered the same stimulus-seeking items that Bem administered. Both were phrased as “To what extent is the following statement true of you:”, and the first statement was “I am easily bored” and the second as “I often enjoy seeing movies I’ve seen before.” Responses were collected on a 5-point scale anchored at 1 (Very Untrue) and 5 (Very True).

Participants then went through a 3-minute relaxation procedure as described in the original paper: people looked at an astronomical photograph while listening to relaxing music. When the 3-minutes had ended, participants clicked a button to acknowledge they were ready, and received instructions about the task. Participants were told:

“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. It is absolutely critical that you focus on only this task and do not perform any other tasks (e.g. check email). When you are ready to begin, please click continue.”

After participants clicked continue they were shown the series of words, each for 3 seconds. As with Bem, the words were drawn from 4 categories: food, animals, occupations, and clothes (see table 1 for a full list of the words). Mirroring 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 randomly assigned 24 words to be practiced; 6 randomly chosen from each of the 4 groups of 12 words. The practice sessions asked people to look at the list of 24 words, and on successive screens, first 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. They could not continue the experiment until they correctly clicked on the appropriate six words and typed the six words in the corresponding boxes.

Table 1

List of Words Used by Category

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

Note—Words are presented alphabetically in this Table, but were presented randomly (across and within categories) to participants.

When the practice session was complete, participants 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 email, 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 email)?". 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".

Results

In order to assess whether or not we observed retroactive facilitation of recall we first had to determine which words were recalled as a function of the ones that were practiced or not. On the surface, this seems like a trivial task, but given that spelling errors were rather prevalent, complete computerized automation could not be used. Instead, we coded the 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 entirely blind as to whether the words were drawn from the practice set or the control set.

Bem (2010) computed a weighted differential recall score (DR) for each participant using the formula:

$$\begin{aligned} \text{DR} = {} & (\text{Recalled Practiced Words - Recalled Control Words}) \times \\ & (\text{Recalled Practice Words + Recalled Control Words}) \end{aligned}$$

In the paper, for descriptive purposes, Bem frequently reports this number as DR%, which is the percentage the score deviated from random chance towards the highest or lowest scores possible (-576 to 576). We conducted the identical analysis on our data and also report DR% (see Table 2). In addition to using the weighted differential recall score, we also report the results from using a simple unweighted recall score, the difference between recalled practice words and recalled control words. For both of these measures, a score of 0 is predicted by random chance, and analysis was conducted using a one-sample t-test.

We did not find any evidence of precognition, as people recalled slightly fewer words from the practice set than from the control set (see Table 2). One concern with the experiment was that it was conducted over the internet and it is unclear the extent to which people fully attended to the key elements of the procedure. We used two methods for excluding these hypothetical inattentive participants. First, we asked them to self-report if they had stopped paying attention at some point during the experiment. Eight people said that they had. Second, we looked at how long people spent on the recall task. Our reasoning was that if people went through that task unusually quickly, it might reflect that they were not particularly focused on the task. The distribution of time was necessarily skewed (i.e., people could take as long as they wanted, but they couldn't go any faster than 0 seconds) so there were no participants who were more than two standard deviations below the mean (which would have reflected negative time on task). We instead used a cutoff of 1 standard deviation below the mean, and this cutoff excluded 7 people from the sample. As can be see in Table 2, neither of these exclusions (either alone or in combination) had any appreciable influence on the effect.

Bem (2010) reported a relationship between sensation seeking and precognitive abilities. He reports a positive correlation across the nine experiments in the paper (*r* = .18) and in Experiment 8 in particular (*r* = .22, *p* = .014). We did not replicate that result. With increases in sensation seeking there was a tiny, and entirely nonsignificant, *decrease* in precognitive ability (as reflected by DR%), *r* = -.063, *p* = .51.

Table 2

Experiment Results

				Weighted Differential Recall		Simple Differential Recall		Percentage of Participants differentially recalling Practice and Control words		
	N	P[1]	C[2]	Mean (DR%)	Statistic[3]	Mean	Statistic	P>C	P = C	P<C
Bem (2010, Study 8) Results	100			2.27%	*t*(99) = 1.92, *p* = .029					
Full Sample	112	8.09	8.43	-1.35%	*t*(111) = -1.31, *p* = .194	-.34	*t*(111) = -1.12, *p* = .265	38.4% (43 of 112)	13.4% (15 of 112)	48.2% (54 of 112)
Removing Self-Identified Inattentive People	104	8.30	8.60	-1.34%	*t*(103) = -1.21, *p* = .230	-.30	*t*(103) = -.93, *p* = .357	39.4% (41 of 104)	13.5% (14 of 104)	47.1% (49 of 104)
Removing People who were too fast on the recall portion (<1SD)	103	8.59	8.99	-1.52%	*t*(102) = -1.36, *p* = .178	-.40	*t*(102) = -1.22, *p* = .226	37.9% (39 of 103)	12.6% (13 of 103)	49.5% (51 of 103)
Removing people from either of those two categories.	95	8.86	9.22	-1.52%	*t*(94) = -1.26, *p* = .212	-.36	*t*(94) = -1.02, *p* = .310	38.9% (37 of 95)	12.6% (12 of 95)	48.4% (46 of 95)

[1] P = the number of practice words correctly recalled (out of 24 possible)

[2] C = the number of control words correctly recalled (out of 24 possible)

[3] Bem uses 1-tailed tests (with good justification) in his paper. Because our replication was not after a specific hypothesis (we were equally open to evidence for precognition and anti-precognition), we report the p-values from two-tailed tests.

Discussion

Why do we not see any evidence of precognition? There are obviously a multitude of possibilities for why we failed to obtain a result similar to Bem, ranging from the mundane (e.g., our sample was more heterogeneous than Bem's) to the exotic (e.g., the quantum mechanics that allow for the detection of future events are also contingent on the specific physical features of the original experiment rooms). For the purposes of this paper we really only care about one possibility: Do we fail to detect precognition because precognition does not exist? In answer to this question we emphatically say, "We don't know. On the one hand, we fail to replicate the effect, but on the other hand, our single failure to replicate is hardly sufficient to seriously undermine an entire paper."

Bem presented nine experiments confirming the existence of precognition. We present one experiment (similarly powered, but burdened by other idiosyncrasies) which seems to show a null effect. In every other psychological domain, that should rightfully be identified as a mild challenge to the original hypothesis, but hardly a severe threat. If we knew for certain that precognition did exist, pure randomness would *frequently* produce a null effect in this experiment, or even the mild reversal we document.

Rather, we interpret our finding as providing exactly the sort of publicly available evidence Bem called for: an effort to scientifically investigate the existence of precognition using a generally valid and agreed upon methodology. His hope, and ours, is that perhaps a handful of researchers will make similar efforts to clarify the effect. Without a doubt, if the effects are real and valid, they would constitute a substantial advance in psychology. Even if that finding feels unlikely[2], its importance would seem to merit the effort of investigation.

For simplicity, we offer two easy subheadings:

What do we claim? That we conducted a very close replication of Bem (2010, Study 8) and failed to obtain a reliable result.

What do we NOT claim? That we have disproven Bem (2010). (We are merely trying to add more data relevant to the question.)[3]

References

Bem D. J. (2010), Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology*, in press.

Hróbjartsson, A. & Gøtzsche, P. C. (2010), Placebo interventions for all clinical conditions. *Cochrane Database of Systematic Reviews 2010*, Issue 1. Art. No.: CD003974.

Geiger, H. & Marsden, E. (1909). On a Diffuse Reflection of the α –Particles. *The Royal Society*, 82 (557), 495-500.

Endnotes

[1] We launched the program once it had been fully prepared by sending an email out to 1,200 people. People started to respond and we left the program running for 41hours, at which point we closed the experiment. This amount of time was arbitrary: we neither stated ahead of time how long we would leave it up, nor did we look at any of the data prior to taking it down. In truth, we saw that there were more than 100 respondents, and decided to close the experiment because one of the authors was getting married the next day and we wanted to be able to talk about it at the reception.

[2] In most papers it seems ridiculous to articulate the "beliefs" of the authors, but somehow that feels relevant here. Indeed, as an example, a powerful observation from meta-analyses of placebo effects has shown that researchers whose investigations seek out placebo effects are substantially more likely to find them than are researchers whose studies are similarly designed, but who are not looking for placebo effects (Hróbjartsson & Gøtzsche, 2010). Placebo effects persist, apparently, especially when the researcher is also in the placebo condition. Accordingly, it is worth noting that the two authors are a bit different from one another in our beliefs. As a way to demonstrate: Both of us lead journal discussion groups at our respective universities and both of us assigned the Bem paper (and highly recommend that you do the same). Both of us were also happy to take bets about the likelihood of replicating one of the studies, and offered odds to our groups of prospective gamblers. One of us offered 4-1 odds against, which, given the vagaries of chance, constitutes fairly substantial positive belief. The other offered 19-1 against, which was designed to capture the .05 alpha level we deal in (while leaving a margin for the house). That can certainly be considered robust skepticism.

Both authors, it should be noted, would have welcomed an outcome that confirmed the original finding. It is said that when Rutherford's lab first documented the existence of the atomic nucleus (Geiger & Marsden, 1909) —in perhaps the most widely described experiment in any scientific discipline— that every physicist immediately went out to replicate the astonishing finding for themselves. Undoubtedly, many were looking to unseat the greatest experimental physicist of the time by undermining the result. But more likely, most just wanted to see it for themselves. That was us. Our lack of a confirmatory finding does not tell us that Bem was wrong, but it sure would have been exciting to confirm that he was right.

[3] If you are interested, you can try being a participant in the study yourself. We have now programmed a slightly different version of the experiment which you can take here: http://consumerbehaviorlab.com/esp1_live/esp1_live.php. At the end of the experiment the program will report back your score (as best judged by a computer scoring algorithm). This algorithm appears to capture (and correct) most misspellings. That should remove the one form of human intervention in the study (and source of bias). The procedure is quite similar to the one reported in this paper, but not quite identical.