Failure to Replicate Retrocausal Recall

Imants Barušs and Vanille Rabier

King's University College at The University of Western Ontario

Author Note

Imants Barušs, Department of Psychology, King's University College at The University of Western Ontario; Vanille Rabier, Department of Psychology, King's University College at The University of Western Ontario.

Vanille Rabier is now at Bimini Biological Field Station, Bahamas.

This study began as Vanille Rabier's undergraduate thesis in psychology at King's University College.

This research was supported by King's University College and Medical Technology (W. B.) Inc. The authors thank Daryl Bem for providing us with the software and instructions for running the experiment; Carolyn van Lier, Jennifer Stewart, and Terra Duchene for interacting with participants; Arwen Sweet for data entry; Sauro Camiletti for advice about statistics; Julia Mossbridge for feedback about the study; and Shannon Foskett for literature searches, critical comments, and proofreading.

Correspondence concerning this article should be addressed to Imants Barušs, Department of Psychology, King's University College at The University of Western Ontario, 266 Epworth Ave., London, Ontario, Canada N6A 2M3. E-mail: baruss@uwo.ca

Abstract

In two temporally inverted memory experiments, Daryl Bem found that participants had better recall for words that were practiced after the recall task than for control words that were not practiced after the recall task. We attempted to replicate the second of Bem's two experiments with the addition of a personality measure. After completing the Six Factor Personality Questionnaire, participants (N = 102) interacted with a computer on which they were shown 48 nouns, one at a time, then asked to type as many of the words as they could recall, and then asked to practice a random selection of 24 of the 48 words. The mean number of practice words that were on the typed recall list was 9.32 whereas the mean number of control words was 9.53 with $t(101) = -0.55$, $p = .58$ (two-tailed), Cohen's $d = -0.060$. The correlation of the difference between practice and control words with Bem's stimulus seeking scale is $r(100) = -.081$, $p = .42$ (two-tailed). Multiple linear regression of Bem's scale on personality facets revealed that Bem's scale does not appear to measure stimulus seeking. The correlation between the difference score and a scale constructed to measure stimulus seeking from the facets of the personality questionnaire was $r(100) = -.063$, $p = .53$ (two-tailed). We found no evidence for retroactive facilitation of recall and no correlations of the difference score with any personality measures. Possible reasons for a failure to find retrocausal recall are discussed.

Keywords: retroactive facilitation of recall, precognition, personality traits, memory

A question that has provoked debate is whether people are able to anticipate future events that they cannot know through ordinary means—a phenomenon called precognition. Precognition can ostensibly take different forms, encompassing precognitive dreaming (e.g., Barušs, 2003; Sherwood & Roe, 2003; Ullman, Krippner, & Vaughan, 1973), presentiment, whereby physiological changes can occur that are consistent with later exposure to stimuli, (e.g., Bierman & Radin, 1999; Mossbridge, Tressoldi, Utts, 2012; Radin, 1997; 2004), and precognitive remote perception (e.g., Dunne, Dobyns, & Intner, 1989; Palmer, 1978; Targ, Schlitz, & Irwin, 2000). A review of precognition research is beyond the scope of this article. Rather, our focus is on a particular experimental protocol introduced by Daryl Bem (2011) as part of his research into the existence of precognition.

Retroactive Facilitation of Recall

As one way of studying precognition, Bem reversed the order of critical tasks in a commonly accepted memory paradigm, whereby words first practiced to commit to memory are more likely to be recalled later than non-practiced words. In his Experiment 8 (out of a series of nine experiments) 100 participants were shown 48 nouns (such as "mechanic" or "pineapple"), one at a time, on a computer screen for three seconds each, and asked to visualize their referents. Then they were given a free recall task in which participants typed as many of the words from the list as they could recall. Next, the computer randomly selected 24 of the 48 words and participants practiced the words by typing them into categories. Experiment 9 was a replication of Experiment 8 with the difference that only 50 individuals participated and, after the free recall task, participants were shown the 24 computer selected words one at a time, organized by category, before typing them into categories. Retrocausal recall would be evidenced by the presence of a greater number of practice than not practice words on the recall list.

As a dependent measure, Bem used $DR\% = [(P - C)(P + C)/576] \times 100$, where P represents the number of practice words on a recall list and C represents the number of control words on a recall list. Bem gave no justification for weighting the difference score by the total number of recalled words beyond saying that each recalled word can be conceptualized as a single trial and that studies with more trials are given more weight in meta-analyses. This method of computation inflates small differences for participants who recall a large number of words, yet it is not at all clear why this would be a desirable outcome. For Experiment 8, the mean *DR%* was "2.27%, $t(99) = 1.92$, $p = .029$, $d = 0.19$" (Bem, 2011, p. 419) and for Experiment 9 it was "4.21%, $t(49) = 2.96$, $p = .002$, $d = 0.42$" (Bem, 2011, p. 420).

Personality Correlates

Bem also created a "stimulus seeking scale" (*SS*) that consisted of two items: "I am easily bored" and the reverse-scored "I often enjoy seeing movies I've seen before" (Bem, 2011, p. 410). There was a correlation of $r = .22$, $p = .014$ between *DR%* and *SS* in Experiment 8, and a correlation of $r = -.10$, $p = .25$ in Experiment 9. Bem attributed the disappearance of the correlation in Experiment 9 to the benefits of the enhanced exercise of the practice words.

Stimulus Seeking is a facet of the personality trait of Extraversion on the NEO Personality Inventory-Revised (NEO PI-R; Bem, 2011). Extraversion has historically been associated with the occurrence of anomalous information transfer (Eysenck, 1967; Palmer, 1978; Rao, 1974). In a meta-analysis, Honorton, Ferrari, and Bem (1998) found a correlation of $r = .21$, $z = 4.57$, $p = .000005$, for 11 studies using free-response tasks in which the measure of extraversion was given to participants before engaging in efforts at extrasensory perception. Nevertheless, the researchers concluded that any correlation of extraversion with forced-choice tasks was an artifact of participants having received the extraversion measure after completing the experimental tasks (Honorton et al., 1998). The conclusion that extraversion is unrelated to forced-choice extrasensory perception tasks has been challenged, however, with the assertion that correlations for forced-choice tasks are comparable to those of free-response tasks under individual testing (Palmer & Carpenter, 1998).

Adopting a broader framework, there appears to have been a decline in the correlation of extraversion with performance on anomalous tasks. Bierman has noted that following the Honorton et al. meta-analysis, the correlation was confirmed with a slightly lower value of $r = 0.18$ in another study, followed by $r = 0.15$ in a subsequent study, and has not only fallen to virtually zero since that time, but become negative in valence (Bierman, 2001). Nonetheless, in the current study, we used a personality measure to seek correlations between extraversion and an increased number of practice words relative to control words on the recall list.

Replication Attempts

Since the publication of Bem's experiments, a number of researchers have attempted to replicate his findings using protocols with varying degrees of conformance to those of Bem. For instance, although some researchers adhered closely to the methodology of the original study (e.g., Cardeña, Marcusson-Clavertz, & Wasmuth, 2009, Ritchie, Wiseman, & French, 2012; Robinson, 2011), other researchers changed the word list and introduced online data collection (e.g., Galak, LeBoeuf, Nelson, & Simmons, 2012). Galak et al. (2012) conducted seven experiments with a total of 3,289 participants. The results of only one of those experiments attained statistical significance with $t(108) = 1.77$, $p = .04$ (one-tailed), $d = 0.17$. Galak et al. also conducted a meta-analysis in which they included Bem's original Experiments 8 and 9 as well as 17 attempts to replicate those experiments. The overall average effect size was *DR%* = 0.04 which was not different from zero. Furthermore, there was no correlation of *DR%* with sensation seeking with $r = -.03$, 95% CI [−0.06, 0.00] (Galak, LeBoeuf, Nelson, & Simmons, 2012).

Hypotheses

The purpose of this experiment was to seek to replicate Bem's Experiment 9 with the addition of a personality test. Given that we initiated this study before the results of any of the replications were available, we tendered the following hypotheses:

1. There will be more practice words than control words on the recall list.
2. Higher scores on extraversion will be correlated with more practice than control words on the recall lists.

Given that we only had Daryl Bem's results at the time we began our study, which were positive for two experiments and had large effect sizes, we expected positive results for our

experiment as well, particularly since we doubled the number of participants Bem needed in order to find statistical significance. However, taking into account the general failure of other researchers to find effects, we ended up using our data to seek to try to determine which parameters could account for the differences in results among studies. We take up this task in the discussion section.

Method

Participants

There were 102 participants in the study (n = 67 females, n = 35 males) with a mean age of 21.2 years of which 67 were women and 35 were men. Thirty-one participants indicated that they were Catholic, 28 indicated that they were Christian, and 22 stated that they had their "Own Beliefs," with smaller numbers endorsing other religious affiliations. Only 17 indicated that they practiced their religion daily or weekly, with 14 saying that they practiced monthly, and 71 that they practiced annually, hardly ever, or never. Participants were recruited either through an on-line sign-up system for students in introductory psychology or directly by the authors or research assistants. Of the 102 participants, 77 received course credit after completing an assignment following their participation and 25 were volunteers who did not receive course credit. Most of the participants had some university education, with nine indicating that they had completed a university degree.

The "optional stopping problem" refers to stopping data collection at the point where the results confirm an experimenter's bias. To avoid this problem, we decided at the outset to collect data from 100 participants, a number based on Bem's work. In fact, we collected data from 102 participants because during the course of the experiment procedural problems appeared to support discarding the data from two participants. Nevertheless, in the end, we included data from all 102 participants in the final sample. Both of the final two participants had equal numbers of practice and control words on their recall lists so that their inclusion in the study had a minimal impact on the results.

Materials

An announcement was placed on the college's computerized sign-up system seeking participants from the participant pool for "Retrocausal Recall Experiment 1" in which participants receiving course credit would be expected to "complete a demographic form, a personality questionnaire and a memory task administered on a computer." In addition, the experimenters and research assistants recruited anyone who found out about the study and was interested in participating, including students in classes without a participant pool. Separate informed consent forms were used for participants who were receiving course credit and those who were not.

A "Demographic Information" sheet was used to collect data regarding age, sex, highest level of education, religious affiliation, and frequency of religious practice.

The paper-and-pencil version of the Six Factor Personality Questionnaire (SFPQ) was used. This empirically developed instrument consists of six factor scales and 18 facet scales, with each facet being measured by six items, each of which is scored on a 5-point Likert scale from "strongly agree" to "strongly disagree." The following are the factors with their facet scales:

Methodicalness consisting of Cognitive Structure, Deliberateness, and Order; Industriousness consisting of Achievement, Endurance, and Seriousness; Independence consisting of Autonomy, Individualism, and Self Reliance; Openness to Experience consisting of Change, Understanding, and Breadth of Interest; Extraversion consisting of Affiliation, Dominance, and Exhibition; and Agreeableness consisting of Abasement, Even-temperedness, and Good-naturedness. Norms were established from the responses of 1067 randomly selected adults from Canada and the United States. Cronbach's alpha ranges from .76 to .86 "for the factor scales and from .55 to .84 for facet scales" (Jackson, Paunonen, & Tremblay, 2000, p. 29).

We used Daryl Bem's computer program for Experiment 9, which included four items requesting information about stimulus seeking, test anxiety, and self-development practices such as meditation; a 3-minute relaxation; and the retroactive facilitation of recall experiment itself. A "Session Record Form," based on one provided by Bem, was filled out by the experimenter, noting whether or not participation was done for course credit, the impression created by the participant on the experimenter, the results, and any comments that the experimenter wished to make.

A "Debriefing Sheet" was provided to all participants upon completion of the study. In addition, a "Research Assignment Questionnaire" was given to students receiving course credit for participation.

Procedure

The study took place in the Psychology Laboratory at King's University College. When participants arrived at the laboratory, they were seated at a desk and provided with a brief verbal description of the study as well as a paper copy of the appropriate consent form to read. After the consent form was signed, participants completed the demographic information sheet and the Six Factor Personality Questionnaire. Then they were escorted to a desk with a computer and told to follow the instructions presented on the computer screen that led them through Daryl Bem's Experiment 9 as described above. At the end of the participants' interactions with the computer, the experimenter wrote down the percentages of practice and control words in the recall list that appeared in the final image on the computer screen and explained them to the participants. Finally, participants were debriefed.

Results

From a total of 1984 words on the participants' recall lists, 106 were not recognized by the computer program as any of the 48 presented words or a common misspelling of any of the words. A research assistant and the first author (IB) judged whether misspelled words should count as being on the list of 48 words, blind as to whether they were practice or control words. The research assistant and first author agreed that 44 of those words should count as being on the list. Of those 44 words, 24 were practice words and 20 were control words. The research assistant and first author did not agree on one word, but IB decided to count it as being on the list. That word turned out to be a control word.

Tests of Hypotheses

Hypothesis 1. A total of 21 more control words than practice words was recalled with a mean of 9.32 (*SD* = 3.69) for practice words and 9.53 (*SD* = 3.30) for control words. A

paired-samples *t*-test was not significant, with $t(101) = -0.55$, $p = .58$ (two-tailed), Cohen's $d = -0.060$, providing no evidence in support of retroactive facilitation of recall.

Hypothesis 2. The subtraction of the number of control words from the number of practice words was designated as *dr*. The correlation of *dr* with Extraversion was $r(100) = -.052$, $p = .60$ (two-tailed). There is no evidence that extraversion plays a role in retrocausal recall. In fact, we obtained no statistically significant correlations of *dr* with any personality factors or facets of the SFPQ.

Stimulus Seeking

The correlation of *dr* with Bem's measure of stimulus seeking was $r(100) = -.081$, $p = .42$ (two-tailed), a finding consistent with Bem's result in Experiment 9. Cronbach's alpha for the stimulus seeking scale was $\alpha = 0.17$ and the correlation of the two scale items was $r(100) = .093$, $p = .35$ (two-tailed). The correlation of *SS* with Extraversion was $r(100) = -.036$, $p = .72$. Multiple linear regression of stimulus seeking on the personality facets provides a solution with two predictors with $R^2 = .14$: $b_1^* = -0.43$ for Breadth of Interest and $b_2^* = 0.32$ for Change. (All multiple linear regression in this study was carried out using stepwise selection with a probability of *F*-to-enter of .05 and a probability of *F*-to-remove of .10.) The zero-order correlations of *SS* with Breadth of Interest and Change were $r(100) = -.26$, $p = .009$ (two-tailed) and $r(100) = .091$, $p = .36$ (two-tailed). These contradictory predictors of inflexibility and adaptability together account for only 14% of the variance in stimulus seeking. The variable *SS* does not appear to be a measure of stimulus seeking.

Although stimulus seeking is a facet of extraversion on the NEO Personality Inventory-Revised, it is not a facet of the Six Factor Personality Questionnaire. However, there are correlations of $r = .31$, $r = .33$, and $r = -.31$ of the NEO PI-R Excitement-Seeking facet with the Dominance, Exhibition, and Seriousness facets of the SFPQ (Jackson, Paunonen, Tremblay, 2000). Thus, a new variable *jss* was created as a sum of the three facet scores (with Seriousness reverse-scored) with a reliability of $\alpha = 0.51$ (where reliability was calculated using facet scores rather than item scores). The correlation of *jss* with *SS* was $r(100) = -.040$, $p = .69$ (two-tailed). However, the correlation of *jss* with Extraversion was $r(100) = .88$, $p < .001$ (two-tailed). Accordingly, *jss* appears to measure"stimulus seeking." The correlation of *jss* with *dr* was $r(100) = -.063$, $p = .53$ (two-tailed) so that there appears to be no significant correlation of stimulus seeking with a greater number of practice than control words on the recall list.

Bem's Dependent Measure

Our dependent measure *dr* had a correlation of $r(100) = .97$ with Daryl Bem's dependent measure *DR%* even though the two equations do not have a linear relationship. Using Bem's measure, *DR%* = −0.20% (*SD* = 13.98%). A one-sample *t*-test against the constant zero was not found to be significant, with $t(101) = -0.14$, $p = .89$ (two-tailed), Cohen's $d = -0.014$. The correlation of *DR%* with Bem's stimulus-seeking scale was $r(100) = -.037$, $p = .71$ (two-tailed) and with *jss* was $r(100) = -.073$, $p = .47$ (two-tailed). A null result was also obtained when Bem's measure was used.

Discussion

A key question is *Why did we fail to find an effect of retrocausal recall whereas Bem did find an effect?* We consider a number of possibilities for failure to replicate Bem's results. The first is that Bem's findings were a statistically rare event, and that the phenomenon of retrocausal facilitation of recall does not occur.

Differences with Bem's Experiment

A second possibility is that retrocausal recall does occur, but that there are critical differences between Bem's experiment and ours that can account for the discrepant results. Two such differences are the instructions provided to participants and the introduction of a personality test as part of the procedure. In Bem's experiment participants were told "'This experiment tests for ESP . . .'" (Bem, 2011, p. 419), whereas in our experiment participants were told that they were going to be doing "a memory task administered on a computer." We did not conceal the nature of the memory task prior to participants' engagement in the experiment (i.e., that we were testing for the effects of future practice effects) and we debriefed participants fully after the experiment. We did not measure participants' understanding of the purpose of the experiment, so we do not know how many participants were unaware of the purpose of the experiment as they engaged in it. Nonetheless, as they were leaving the laboratory, it was clear that some of the participants were unaware that we had been seeking evidence of future influence on a task that had been completed in the past. It is possible that understanding that one is engaged in a precognitive task could somehow facilitate performance on such a task. On the other hand, such a demand could create anticipatory anxiety and attenuate any precognitive performance. Future studies could manipulate participants' expectations regarding the research to determine whether expectations can account for the discrepancies among investigations.

The second main difference between our study and Bem's is that we administered a personality test prior to the recall experiment whereas Bem did not. It is possible that responding to the personality test somehow suppressed participants' abilities to benefit from practice with 24 of the words after writing the recall list. Indeed, any stimulus could serve as a prime of one kind or another. Even signing one's name, as required for informed consent, can affect a participant's performance on a subsequent task (Kettle, 2011). However, as discussed in the introduction, many researchers have administered personality tests prior to tasks involving anomalous information transfer without such tests compromising performance to the point where the anomalous effects were vitiated.

It is also possible that Bem's participants were more conscientious in their engagement with the task compared with our participants and that a high level of conscientiousness and task focus are required for success (cf. Franklin & Schooler, 2013). That degree of engagement in the task could be a factor is suggested by a correlation between *dr* and Industriousness of $r(100) = .14$, $p = .16$ (two-tailed). The correlation of *DR%* with Industriousness was $r(100) = .17$, $p = .09$ (two-tailed). It is possible that those who were more focussed on the task were more likely to demonstrate retrocausal recall. However, those correlations were not statistically significant. Also, item two of the Session Record Form required the experimenter to rate whether or not the

participant was taking the experiment seriously. All but one of the participants received a rating of "yes." This is a crude measure but, together with the Industriousness scale, the only measures we had of task engagement. Thus, although it remains a possibility, there is no empirical support for the idea that our participants were less conscientious in their approach to the experimental task than Bem's participants. However, this is a variable that researchers might wish to measure in future studies.

Psi-Missing

A third possibility is that "psi-hitting" was counterbalanced by "psi-missing." "Psi-hitting" refers to detecting and writing down practice words, and "psi-missing" refers to detecting and avoiding writing down practice words. Is there any way to detect whether psi-missing is occurring given the data that we have? Here is one way to think about it. If an approximately fixed proportion of practice words recalled results from the anomalous cognition of practice words, then the right tail of the distribution of practice words on the recall list should be "pushed out" even if there is no mean shift of the distribution relative to control words. In fact the distribution of practice words had a positive skew of $g_1 = 0.23$ although it was not statistically significant ($z = .95$, $p = .17$, one-tailed). If psi-missing were to not just be a matter of passively avoiding practice words, but "proving" that precognition is impossible, then participants who are psi-missing, might be expected to anomalously draw on the list of control words to which they had been exposed in such a way as to inflate the list of control words. Note that this would be in addition to any possible anomalous access by all participants to the initial list of words. The skew for control words is only $g_1 = 0.06$ ($z = .23$, $p = .41$, one-tailed). On the basis of this reasoning, these results are consistent with psi-hitting but not active psi-missing.

If psi-missing and psi-hitting were both to be present, it might be possible to distinguish between them using some combination of the personality facets. In other words, we are looking for a synthesized personality measure with the property that those with high scores on that personality measure would have high positive scores on *dr* signifying psi-hitting, whereas those with low scores on the derived personality measure would have high negative scores on *dr* signifying psi-missing. Constructing such a measure would not mean much in and of itself. It would need to be tested on a second sample. We did this by splitting the sample into two random samples of 51 participants each. One half of the sample was used to create new personality variables p_1, p_2, and p_3 that were the sums of facets oriented so as to have positive correlations with *dr*. For instance, the new variable p_1 was the sum of two facets that had correlations of at least $r(49) = .20$ with *dr*. We obtained a correlation of $r(49) = .36$, $p = .01$ (two-tailed) of the variable p_1 with *dr*, which suggested that p_1 was a personality trait that could distinguish psi-missing from psi-hitting. However, the correlation of p_1 with *dr* for the second half of the sample was $r(49) = -.10$, $p = .51$ (two-tailed) indicating a failure to replicate the findings from the first half of the sample. The same pattern of success and failure occurred for the other new variables p_2, and p_3. There is no evidence that the personality factors and facets can be used to identify those who can succeed at the retrocausal recall task. Thus, we found no evidence to support the hypothesis that psi-missing is cancelling psi-hitting.

Decline Effect

A fourth possibility is that our experiment suffered from the decline effect. The decline effect refers to a decline in performance by participants over the course of an experiment seeking to demonstrate anomalous phenomena (Bierman, 2001). The cumulative deviation is the sum of deviations from zero as a function of a participant's position number in the temporal sequence of participants. In our study the cumulative deviation reached a maximum value of plus 17 words with participant 22, and a minimum value of minus 41 words with participant 69 before returning to minus 21 with participant 100. Cumulative deviation as a cubic function of participant position number has good fit with $R^2 = .68$. This is not surprising, given that we noticed a cubic relationship and then fitted a cubic equation to it. A linear equation gives $r(100) = -.72$, $p < .0005$ and $R^2 = .51$, so there is an overall decline (cf. Cardeña, Marcusson-Clavertz, & Wasmuth, 2009). However, the correlation of participant position number with *dr* only gives $r(100) = -.004$, $p = .97$ (two-tailed). Also, analyzing the data using a paired samples t-test at the maximum cumulative deviation of 22 participants gives $t(21) = 1.15$, $p = .26$ (two-tailed). In other words, even selectively analyzing only the data before the decline does not allow us to reject the null hypothesis. Thus, although these results roughly follow a decline pattern there is no evidence in this case that such a pattern is anything other than stochastic wobble.

The decline effect can also refer to a decline in effect size with the temporal sequence of similar experiments, as discussed previously in the case of correlations of anomalous events with extraversion (Bierman, 2001). This could be due to ordinary processes, such as selective reporting, or anomalous ones, although it is not clear exactly how those would work. Daryl Bem proposed the retroactive facilitation of recall protocol and obtained significant results but, presumably as a result of the decline effect, researchers trying to do so afterwards have been unable to replicate the results. It is not clear what counts as a replication, however. Using a somewhat different retrocausal recall protocol, Traxler, Foss, Podali, and Zirnstein (2012) timed participants reading a 1,009 word text and then had them read that text again or read a different text. The time that it took to read the text for the first time should have been quicker in the condition where the same text was read again rather than the condition in which a different text was read, but there was no difference between conditions. If this experiment were to be considered a unique variation on precognition experiments, on the order of variation of Bem's retrocausal recall experiments, then the prediction would be that an effect would initially be present for Traxler et al. before declining. However, there was no effect to begin with in the retrocausal text reading study (even though one would reasonably be expected) before declining. In other words, the criteria for the occurrence of a decline effect have been insufficiently established to serve as a tenable explanation for the results of our study.

Experimenter Effect

A fifth possible explanation is that the "experimenter effect" suppressed any retroactive facilitation of recall. There are two possible mechanisms for the experimenter effect: psychological and anomalous. An experimenter could influence participants through ordinary psychological processes (e.g., by creating subtle cues that create demand characteristics) in such a way as to improve their performance (Braude, 2002; Smith, 2003b). In addition to the two authors, three research assistants interacted with participants. Differences in *dr* did not vary as a function of

the experimenter, $F(4) = .61$, $p = .66$ (two-tailed), $R^2 = .024$. If experimenters suppressed the effects through ordinary psychological processes, then all were equally effective.

It is possible that the overall ambient tone of participants' interaction with experimenters in our laboratory differed from that at Cornell University where Bem's studies were conducted. However, not only does our laboratory have a pleasant décor, but the ambiance was intended to promote relaxation by having stuffed animals on top of a filing cabinet and by using indirect lighting while participants interacted with the computer.

The Session Record Form contained seven items assessing an experimenter's impressions of each participant. Taken together, these seven items had a reliability of $\alpha = .77$ and face validity, with item content tapping the positivity of participants' experiences during the experiment. A new variable *obs* was created by adding together the seven items. The values of *obs* varied from 11 to 20 with a possible range of 7 to 21, had a median of 16, and a mean of 16.20 ($SD = 2.50$). Although it would be interesting to compare these values with comparable measures obtained by Bem, they were not reported for his study. On the face of it, the high median and mean scores do not support the hypothesis that the laboratory environment compromised participant performance. Moreover, the correlation of *obs* with *dr* was not significant, $r(100) = .026$, $p = .80$ (two-tailed), further contradicting the notion that the quality of participants' experiences in the laboratory contributed to our findings.

Another possibility is that the experimenters affected the outcome of the study through some anomalous yet unspecified mechanism (Delanoy, 1996; Smith, 2003a; 2003b; Watt & Ramakers, 2003; cf. Kennedy, 2000). For instance, any actual ability of participants to "perceive" words from the practice list at the time of recall could be precisely compensated by the experimenters' nonconscious ability to interfere with such perception. Alternatively, experimenters could affect a computer's pseudo-random number generator so as to ensure that the lists of practice words had about half of the words from recall lists on them. Perhaps participants correctly perceived the words that they were to practice, but the experimenters somehow anomalously changed that list so that those words ended up being control words rather than practice words. However, speculations such as these raise the question of the potential role of remote influencing in scientific research.

Remote Influencing

Remote influencing refers to a person's ability to affect a biological process or physical event without any known physical interaction with the process or event. As an extension of the experimenter effect, a sixth possibility is that participants, either alone or in nonconscious collaboration with experimenters, exerted an influence on the computer's pseudo-random number generator so as to create the list of practice words in the first place. That could be viewed as a real-life application of the Kochen and Specker Theorem in quantum mechanics, whereby observable variables do not have values until such time as a decision is made as to what to measure, or those values change depending upon what else is being measured (Kochen & Specker, 1967; Barušs, 2008; 2009). In such a scenario, Daryl Bem's participants affected the pseudo-random number generator so as to load the practice list with a preponderance of words from their recall list and participants in our study loaded the practice list with only about half of the words from the recall list. To complicate matters, it is not only the participants and experimenters

who could be contributing to the results of a study but anyone who eventually comes in contact with the data or who reads the final paper (Bierman, 2001). Although these are speculative possibilities, our study provides no evidence to support the experimenter effect or remote influencing more generally. Those are variables that could be explicitly explored in further research.

Conclusion

In this experiment we failed to replicate Bem's findings and found no support for retrocausal recall. One possibility is that we have not identified procedural or individual difference variables that mediate the phenomenon of retrocausal recall and that are associated with its manifestation in a laboratory. An alternate possibility is that retrocausal recall does not occur, or at least, that it cannot be demonstrated in a psychological laboratory using introductory psychology students as participants.

Daryl Bem is part of a group of researchers led by Patrizio Tressoldi who have meta-analysed 82 precognition experiments that "became available" since the year 2000 (Tressoldi, Rabeyron, Duggan, & Bem, 2013, p. 9). Their analyses revealed robust evidence for various forms of precognition with the exception of the retroactive facilitation of recall. For the retroactive facilitation of recall experiments they found a split between exact and inexact replications with 14 exact replications having an overall effect size of $d = 0.09$ and 11 inexact replications having an effect size of $d = -0.001$. Much of the data in the inexact experiments came from a large online sample in which there was no monitoring of participants' behavior during the course of their participation (Tressoldi, et al., 2013). Based on these data, the effect size in our experiment should have been around $d = 0.09$ rather than the $d = -0.014$ that we did find (using Bem's dependent measure) and which is numerically larger in the reverse direction than the negative effect size for the inexact replications.

Tressoldi et al. speculated that precognition experiments were sensitive to the "fast thinking" vs. "slow thinking" distinction made by Daniel Kahneman (2011). Fast thinking, which involves implicit cognition without explicit rational processing (such as in presentiment studies using galvanic skin response measures to differentially emotionally valent stimuli) is conducive to the manifestation of precognition. Slow thinking, which requires sustained explicit cognition over the span of minutes (such as in the retrocausal recall experiments) is not conducive to the manifestation of precognition (Tressoldi, et al., 2013). However, given that other laboratory-based retrocausal recall studies have been successful, slow thinking might just substantially attenuate precognition rather than eliminate it altogether. Nonetheless, this is a variable that could be explicitly investigated in future studies.

Cardeña, Marcusson-Clavertz, and Wasmuth (2009) have pointed out that anomalous information transfer usually occurs when stimuli are emotionally charged and have recommended that an emotional component be included in future retrocausal recall experiments (Cardeña, Marcusson-Clavertz, & Wasmuth, 2009). In their meta-analysis Tressoldi et al. found that the most successful experiments were those that required the "precognitive detection of erotic stimuli" (Tressoldi, et al., 2013) apparently supporting the importance of emotionality. It would be fairly straigthforward to set up protocols in which words paired with emotionally charged images could be used as stimuli to determine whether emotionality plays a role in retrocausal recall.

There has been considerable interest among psychologists in Daryl Bem's precognition

experiments, perhaps because they challenge conventional ideas that we have about the nature of reality. We think that it is important to remember, in that regard, that science is a strategy for acquiring knowledge that is based on empirical investigation and the resolution of discrepancies through the identification of relevant parameters. In science, ideas about reality are based on the results of scientific exploration and discarded when they have poor goodness-of-fit to the data (Barušs, 1996). We feel that it is important to continue this line of investigation to seek to resolve the discrepancies between experiments to determine which theories are the most plausible.

References

Barušs, I. (1996). *Authentic knowing: The convergence of science and spiritual aspiration*. West Lafayette, Indiana: Purdue University Press.

Barušs, I. (2003). *Alterations of consciousness: An empirical analysis for social scientists.* Washington, DC: American Psychological Association.

Barušs, I. (2008). Characteristics of consciousness in collapse-type quantum mind theories. *Journal of Mind and Behavior, 29*(3), 255–265.

Barušs, I. (2009). Speculations about the direct effects of intention on physical manifestation. *Journal of Cosmology, 3*, 590–599.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407–425.

Bierman, D. J. (2001) On the nature of anomalous phenomena: Another reality between the world of subjective consciousness and the objective world of physics? In P. Van Loocke (Ed.), *The physical nature of consciousness* (pp. 269–292). Amsterdam, Netherlands: John Benjamins.

Bierman, D. J., & Radin, D. (1999). Conscious and anomalous nonconscious emotional processes: A reversal of the arrow of time? In S. R. Hameroff, A. W. Kaszniak, & D. J. Chalmers (Eds.), *Toward a Science of Consciousness III: The third Tucson discussions and debates* (pp. 367–385). Cambridge, MA: MIT Press.

Braude, S. (2002). *ESP and psychokinesis: A philosophical examination* (Revised ed.). Parkland, FL: Brown Walker Press.

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

Delanoy, D. L. (1996). Experimental evidence suggestive of anomalous consciousness interactions. In D. N. Ghista (Ed.), *Biomedical and life sciences: Proceedings of the Second Gauss Symposium, August 2–8, 1993, Munich* (pp. 398–410). Braunschweig/Wiesbaden, Germany: Vieweg.

Dunne, B. J., Dobyns, Y. H., & Intner, S. M. (1989). Precognitive remote perception III: Complete binary data base with analytical refinements (Technical Note PEAR 89002). Princeton, NJ: Princeton University.

Eysenck, H. J. (1967). Personality and extra-sensory perception. *Journal of the Society for Psychical Research, 44*(732), 55–71.

Franklin, M. S. & Schooler, J. W. (2013, June). *Can practice effects extend backwards in time? An overview of seven years of experimentation*. Paper presented at the 32nd annual meeting of the Society for Scientific Exploration, Dearborn, Michigan, USA.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology, 103*(6), 933–948.

Honorton, C., Ferrari, D. C., & Bem, D. J. (1998). Extraversion and ESP performance: A meta-analysis and a new confirmation. *The Journal of Parapsychology, 62*(3), 255–276.

Jackson, D. N., Paunonen, S. V., & Tremblay, P. F. (2000). *Six Factor Personality Questionnaire technical manual*. London, Ontario, Canada: Research Psychologists Press.

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

Kennedy, J. E. (2000). Do people guide psi or does psi guide people? Evidence and implications

from life and lab. *The Journal of the American Society for Psychical Research, 94*, 130–150.

Kettle, K. L. (2011). *The general self-concept* (Doctoral dissertation). Retrieved from ProQuest, UMI Dissertations Publishing. (NR89811)

Kochen, S. & Specker, E. P. (1967). The problem of hidden variables in quantum mechanics. *Journal of Mathematics and Mechanics, 17*(1), 59–87.

Mossbridge, J., Tressoldi, P., & Utts, J. (2012). Predictive physiological anticipation preceding seemingly unpredictable stimuli: A meta-analysis. *Frontiers in Psychology: Perception Science, 3*, 1–18.

Palmer, J. (1978). Extrasensory perception: Research findings. In S. Krippner (Ed.) *Advances in parapsychological research: Volume 2: Extrasensory perception* (59–243). New York, NY: Plenum Press.

Palmer, J. & Carpenter, J. C. (1998). Comments on the extraversion–ESP meta-analysis by Honorton, Ferrari, and Bem. *The Journal of Parapsychology, 62*(3), 277–282.

Radin, D. I. (1997). Unconscious perception of future emotions: An experiment in presentiment. *Journal of Scientific Exploration, 11*(2), 163–180.

Radin, D. I. (2004). Electrodermal presentiments of future emotions. *Journal of Scientific Exploration, 18*(2), 253–273.

Rao, K. R. (1974). Psi and personality. In J. Beloff (Ed.), *New directions in parapsychology* (pp. 60–76). London, England: Elek Science.

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), 1–5.

Robinson, E. (2011). Not feeling the future: A failed replication of retroactive facilitation of memory recall. *Journal of the Society for Psychical Research, 75*(904), 142–147.

Sherwood, S. J. & Roe, C. A. (2003). A review of dream ESP studies conducted since the Maimonides dream ESP programme. *Journal of Consciousness Studies, 10*(6–7), 85–109.

Smith, M. D. (2003a). The psychology of the "psi-conducive" experimenter: Personality, attitudes towards psi, and personal psi experience. *The Journal of Parapsychology, 67*(1), 117–128.

Smith, M. D. (2003b). The role of the experimenter in parapsychological research. *Journal of Consciousness Studies, 10*(6–7), 69–84.

Targ, E., Schlitz, M., & Irwin, H. J. (2000). Psi-related experiences. In E. Cardeña, S. J. Lynn, & S. Krippner (Eds.), *Varieties of anomalous experience: Examining the scientific evidence* (pp. 219–252). Washington, DC: American Psychological Association.

Traxler, M. J., Foss, D. J., Podali, R., Zirnstein, M. (2012). Feeling the past: The absence of experimental evidence for anomalous retroactive influences on text processing. *Memory and Cognition, 40*, 1366–1372.

Tressoldi, P., Rabeyron, T., Duggan, M., & Bem, D. J. (2013). *Feeling the future: A meta-analysis of 82 experiments on the anomalous anticipation of random future events.* Manuscript submitted for publication.

Ullman, M., & Krippner, S. (with Vaughan, A.). (1973). *Dream telepathy*. New York, NY: Macmillan.

Watt, C. & Ramakers, P. (2003). Experimenter effects with a remote facilitation of attention focusing task: A study with multiple believer and disbeliever experimenters. *The Journal of Parapsychology, 67*, 99–116.