INFOMPL - Research report
Foreseeing the future in a recall task: a replication study

Cándido Otero (6786170), Alissa Hoevenaars (4294661)

Utrecht University

c.oteromoreira@students.uu.nl, a.hoevenaars@students.uu.nl

Abstract

Normally, studies concerning facilitation of recall test whether words that are practiced prior to a recall task are more often remembered compared with control words. However, the 2011 study of Bem [1] found that practicing words after a recall task also results in a higher ratio of recalled practice words compared with control words, indicating that participants had the ability to "feel" the future. This remarkable finding has led to many replication studies [2][3][4][5] with inconsistent findings. To get more clarity on this topic, our study replicated one of Bem's experiments. 43 participants took part in our online experiment where subjects were presented with 48 words, a recall task, and two practice tasks for which 24 words were randomly selected. Multiple t-tests and mixed-effects models were used to investigate if practice words were more often remembered than control words and if stimulus seeking had an effect on this. Furthermore the effect of several other factors, such as word category, word frequency, and word length were examined. Only the word category was found to have an effect on the recall of words. The results were not in accordance with the finding of the original study, and do not indicate evidence that humans can foresee the future.

Index Terms: psi, precognition, experimental psychology, facilitation of recall, replication

1. Introduction

It could be very useful to sense future events. Imagine, for example, that you are about to buy a lottery ticket and you know in advance which ticket would get you a prize. Or, imagine that you are aware that a terrible accident is going to happen on your way to work, and you could decide to stay home. This ability would give you the opportunity to ban disappointment and horror while you can choose for security and pleasure. It may sound like something out of a sci-fi book. However, experimental psychologists have recently taken a related discussion very seriously: the existence of psi.

Psi is concerned with unusual forms of information processing that cannot be explained by physical or biological mechanisms [1]. Some examples of forms of psi are transferring thoughts to another person in the absence of any recognized form of communication (telepathy), influencing physical systems without physical interaction (psychokinesis), and the ability to be aware of, or emotionally affected by, future events without prior knowledge (precognition and premonition, respectively).

Although there have been some studies that indicate the existence of psi, psi is generally not accepted among researchers. Previous studies with results in favor of psi have found that people are able to anticipate future events [6][7]. However, other studies did not find evidence of this phenomenon [2][3][4]. Apart from the contradicting findings, there is currently no explanation for the existence of psi and it contradicts the general beliefs about time, causality, and reality. As a consequence, a large amount of researchers consider psi to be impossible [8]. However, psi has received quite some attention in the field of empirical psychology in the last ten years, mostly from researchers who want to shed some light on controversial claims made by one of their colleagues.

In 2011, Bem [1] published a paper in which he investigated precognition and premonition. In the study, nine experiments were conducted to investigate whether well-known psychological effects also can be observed in the reverse order. One example of a psychological effect that Bem investigated was the occurrence of avoidance behavior after the presentation of a fearful stimulus (e.g., a picture of a snake). A regular study to investigate the effect of avoidance behavior would first introduce a frightening stimulus, and then, measure the behavioral response after the presentation of this unpleasant stimulus, to examine the effect of the stimulus on the behavior. Bem, however, was interested in whether there was a behavioral response before the introduction of the stimulus. Following this idea, he investigated if people were (subconsciously) aware and affected of the presence of future stimuli.

The study of Bem [1] showed promising results for the existence of psi. Out of the nine experiments that were conducted, eight were found to produce a significant effect. The significant p-values ranged from 0.002 to 0.039 and the mean effect size (Cohen's d) across all experiments was 0.22. Furthermore, stimulus seeking (which was defined as susceptibility to boredom and tendency to seek out stimulation) was found to positively correlate with psi performance in five out of the nine experiments. People that scored higher on stimulus seeking showed a higher psi performance.

The results might seem promising at first sight. However, the study of Bem [1] has caused a lot of concern among other researchers. It has even been stated that Bem's paper started a revolution in experimental psychology [9] because it unveiled bad research practices within the field and the demand for replications. Bem has been accused of extensive pilot testing [10], optional stopping [10], and selective reporting [2], which can all lead to erroneous significant results.

A good way to determine whether the Bem's results were found by chance is by replicating the experiment. Multiple replication studies have not found significant results [2][3][4]. However, a meta-analysis[5] in which the results of 90 experiments were examined, found decisive evidence for an effect of stimuli before they were presented. Interestingly, the meta-analysis was co-written by Bem himself. So, the debate about Bem's study and the existence of psi is not over yet.

To do a first-hand analysis on this controversial topic, the current study will replicate one of the experiments of the study of Bem [1]. The experiment that is replicated is experiment 9, which was used to investigate facilitation of recall. This effect refers to the improved performance on memory tasks when a target element is repeatedly displayed before a recall task. In the study of Bem, on the other hand, the recall task takes place before the target stimuli are presented again. The results of this experiment revealed that words that were randomly selected for the practice task were more often mentioned in the recall task compared with the words that were not selected for the practice task ($t(49)= 2.96$, $p = .002$, $d = 0.42$). Stimulus seeking was not correlated with psi performance, although this was the case in a version of the experiment with one practice exercise less (experiment 8).

The current study is very similar to Bem's study [1], although there are some differences. First of all, different words were used in the task of this study because the familiarity of the words in the original study varied. Second, the current study was conducted online instead of in-person. Finally, we use a more complex analysis compared with the original study. Nevertheless, most aspects of the experiment remained the same (e.g., number of words, word categories, and the tasks). This allows for a fair comparison between our own and the original results and helps to shed more light on the question if people are affected by future.

2. Methods

2.1. Participants

Data were collected from 43 participants. Three of the participants were excluded due to abnormal performance, as described in section 2.6. Therefore, data from 40 participants (25 female and 15 male) were used for the analysis. The age of the participants ranged from 21 and 70 years (M = 31.68 years, SD = 11.08 years). Participants were recruited among our own acquaintances. There were no requirements to participate in the experiment, apart from the requisite that participants had to be familiar with the English language.

2.2. Materials

Due to the COVID-19 pandemic, it was not possible to conduct the experiment on a physical location, as in the original experimental setup. To circumvent this problem, we created an online experiment based on Ibex (Internet Based EXperiments)[1], a tool written in JavaScript to run online psycholinguistic experiments created by Alex Drummond. Instead of using a vanilla version of this software, we used PCIbex (PennController for Internet Based Experiments)[2][11], which is a library that provides a number of additional controllers to allow for a more flexible implementation of the experiments and an improved user experience out-of-the-box.

Our experiment was hosted on PCIbex Farm[3] where participants could freely access and partake in the experiment in a completely remote fashion. An online experiment has the advantage that there is no possible interference from the experimenter's side, and that the experiment can not vary between participants (e.g., it is not possible to ask additional questions to one participant, but not to another). Furthermore, participating in an online experiment is less time-consuming which can potentially make it easier to recruit a reasonable amount of participants given the scope of this research study.

We asked people to participate in the experiment on a computer to ensure that tasks were fully visible without having to scroll down the window. Also, doing it on a phone was strongly discouraged, as people are more likely to get notifications and lose focus on the experiment. The experiment runs on any modern web browser that can execute JavaScript. To facilitate replication of our experiment, we made a a git repository available with our standalone PCIbex setup, along with the gathered data and the necessary script to parse it[4].

In his paper [1], Bem encourages the replication of his experiments. He even invites other researchers to contact him so he can send them software to carry out their experiments and detailed instructions. In order to replicate the study as closely as possible, including the directions given to the subjects at each task, we contacted Bem to get access to his instructions. Unfortunately, we did not get a reply. Nevertheless, we used some of the directions

[1] Original Ibex: `https://ibex.spellout.net`

[2] PCIbex library: `https://doc.pcibex.net`

[3] Our experiment: `https://farm.pcibex.net/r/XPIuJP` (**note**: this is a demonstration link, responses are no longer being collected. You can create a copy of our experiment and run it on your own PCIbex Farm by clicking on the top banner)

[4] Replication material: `https://github.com/omcandido/test_bem`

he quoted in his original paper to stay as close as possible to the original instructions he gave to his participants.

One aspect that we changed about the original experiment is the words that were used. In the original experiment, there were both frequent and rare words (for instance, "apple" has a frequency of 61094 while "bricklayer" has a frequency of 160). For this experiment we decided to use words within a smaller range of frequencies to minimise the possible interfere of the word frequency on the recall task [12]. To this end, we picked 48 words with frequencies between 1000 and 50000 sampled from the Corpus of Contemporary American English[5]. These thresholds seemed appropriate to ensure that participants were familiar with the words even if they were not native English speakers, while avoiding trivial words that could be simply guessed instead of recalled. For each of the four categories (animals, clothing, foods, and occupations), we selected 12 words, as in the original design. The word length was not controlled for. The full list of words with their corresponding frequency can be found in Appendix A.

2.3. Design

A within-subjects design was used to investigate if practice words were more often recalled than control words. Although the practice words differed from person to person, the tasks were identical for everyone. There was no between-subjects factor.

2.4. Measurements

Differential recall score
Identical to the original experiment [1], psi performance was measured with a weighted differential recall (DR) score. The formula that was used is:

$$DR\% = [(P - C) \times (P + C)]/576 \qquad (1)$$

P stands for the number of recalled practice words, whereas C defines the amount of recalled control words. (P – C) was multiplied by the participant's overall recall score (P + C) to give more weight to participants who recalled more words. Furthermore, (P – C) x (P + C) was divided by 576 to express the DR-score as a percentage of the maximum possible DR-score (i.e., the result of recalling all practice words but no control words). The DR% can range from -100% to 100%. A negative DR% means that less practice words were recalled compared with control words. On the other hand, a positive DR% score could indicate the existence of psi because participants with a DR% above zero were able to recall more practice words compared with control words.

Stimulus seeking score
The second measurement that was replicated from the original analysis was the stimulus seeking (SS) score. This score was obtained through two questions: (1)*"I am easily bored"* and (2)*"I often enjoy seeing movies I've seen before"*, to which subjects can reply using a 5-level scale response ranging from *"Very true"* to *"Very untrue"*, corresponding to a score of 5.0 and 1.0, respectively. It is important to note that the scale of the question *"I often enjoy seeing movies I've seen before"* is reversed before the calculation of the SS-score (e.g, a score of 4.0 will be changed to a 2.0). Whereas answering *"Very true"* to *"I am easily bored"* would indicate the highest degree of stimulus seeking, the same answer to the question *"I often enjoy seeing movies I've seen before"* would indicate the lowest degree. After the reversal of the score for the movie-question, the SS-score for each participant was calculated as a value between 1.0 and 5.0 by averaging the two responses.

2.5. Procedure

Participants were asked to partake in an online experiment. The experiment started with a few instructions. Participants were asked to use a laptop or a desktop computer and make sure they would not get interrupted and stay focused during the experiment. Furthermore, participants were informed about the length of the experiment, that data were collected anonymously, and that they could stop their participation at any point. Participants that chose to continue, were asked to fill in their age and gender, after which they were presented the two SS questions. This was followed by an instruction screen that asked participants to relax. When the participant was ready, they could choose to start the experiment.

Participants were informed that 48 different words would be displayed in intervals of 3 seconds, and were asked to visualize each word during its display (for example, if they see the word "*mountain*", they are asked to visualize a "*mountain*"). The words were displayed in random order (see Appendix A for a full list of the words). After the presentation of all words, participants were given a (surprise) free recall test in which they were asked to write down as many words as they could remember from the previously presented words. Then, 24 of the 48 words were randomly selected as practice words for further use in the experiment. We also refer to these words as target words. The remaining 24 words were not presented again during the experiment, and served as control words.

The experiment continued with two practice exercises. First, participants were informed that they would be presented with 24 words they saw earlier and that the words would be presented by category (foods, animals, occupations, or clothing). Again, they were asked to visualize the words. Next, followed another practice task. All 24 practice words were presented simultaneously randomly distributed along a 4x6 grid. The participants were

[5] https://www.english-corpora.org/coca/

asked to click on all 6 words that belonged to a certain category. If the word belonged to the right category, it would turn green. Otherwise, and warning message was shown. After selecting all the category words, participants had to rewrite these 6 words. If people made a typing error, they were made aware of this with a message. All words had to be written correctly before continuation of the experiment was possible. This task was repeated for all categories and the location of the words in the grid was reshuffled at each iteration.

After the completion of the experiment, participants were thanked for their participation and informed that the results from the tasks they just completed would be used to study memory and how different sets of words affect recall in individuals. In order to preserve the internal validity of the experiment, they were asked to not talk about the experiment to other people who also were planning to participate in this study. The experiment took around 15 minutes to complete.

2.6. Data preprocessing

Before proceeding to analyze the data, the results had to be parsed into consumable data frames. The data were downloaded in CSV format and imported into R [13].

The first step was to transform the recalled words entered by the users into lowercase and trim trailing and leading whitespaces. This made it possible to compare the words that were entered by the subjects with our 48 vocabulary words. In order to automatically correct as many typing errors as possible, we tested for word similarity by using the Levenshtein distance. Each word that did not have an exact match with one of the 48 terms of the vocabulary, was replaced with its closest match if there was a unique vocabulary term that had a Levenshtein distance of 1. This threshold of 1 was used because the smallest Levenshtein distance between two words in the vocabulary was 2 (when computing Levenshtein distance, we always use a cost of 1 for insertion, deletion, and substitution). Thanks to this method, 25 typos were automatically corrected. Many of theses errors were plurals of the target words (e,g., *cereals* instead of *cereal*). Additionally, 5 words had to be manually corrected (e.g. *pingoen* instead of *penguin*, where the Levenshtein distance is 3). In total, 11 words were discarded as they were not part of the vocabulary set. The relatively low ratio of typos suggests that using non-native English speakers as participants was not a problem for the recall task.

Then, we calculated the number of unique words that each participant recalled correctly (regardless of it being a control or a target word), see Figure 1. This was done as a first step to detect outliers. Excluding extreme outcomes is important because these values can have a major misleading influence on the results. For example, participants who recalled very few words might not have been paying enough attention to the task (or simply were not interesting for our study). Similarly, participants who recalled many words (e.g., one participant suspiciously got 47 out of 48 words correct), might have not used their memory alone to recall the words.

To detect the outliers, we used the same criterion as another recent recall study [14] based on the interquartile range (IQR). All responses that laid within 1.5 times IQR from the mean number of correctly recalled words were included in the analysis. According to this method, participants that recalled fewer than 4.99 or more than 30.49 words correctly had to be excluded for the analysis. These limits resulted in the the removal of three participants, because of the atypical number of recalled words (2, 33 and 47 words).

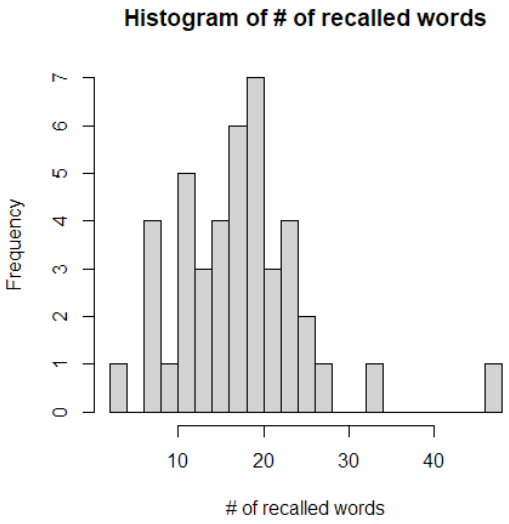

Figure 1: *Histogram showing that the number of recalled words per participant follows a normal distribution. Subjects that recalled less than 5 or more than 30 words are considered outliers.*

Once all outliers were removed, and the stimulus seeking score was calculated (see section 2.4), two different data frames were created. The first data frame contained one row per participant with information about each participant and their DR- and SS-score, see Appendix B. This data frame was intended to replicate the original analysis of Bem's study. The second one was comprised of 48 rows for each of the 40 participants (one per vocabulary term, 1920 rows in total) and contained information about the word (frequency and category), whether the user had recalled the word or not, and whether it was a target or a control word for that given subject. The purpose of this data frame was to perform some exploratory analysis on the data to determine what factors might influence the recall on words.

	Full sample				High SS			Low SS		
Experiment	p	d	df	Correlation with SS	p	d	df	p	d	df
Original	0.002	0.42	49	-0.10 (p = 0.25)	0.049	0.44	15	0.013	0.40	33
Replication	0.975	0.01	39	-0.13 (p = 0.41)	0.237	-0.27	19	0.200	0.30	19

Table 1: *Comparison between the results of the original experiment and our results. The effect of the target words on the psi performance was examined with a t-test that tested if the DR-score differed from μ=0.*

2.7. Analysis

To investigate the existence of psi, the performance of the participants was examined with two different analysis methods using R.

For a fair comparison with the experiment we are replicating, three two-tailed t-tests were conducted to compare our results to the results of the original study. Each t-test was used to investigate if the DR-score significantly differed from 0. For each of the three t-tests we used a different sample. One analysis included all participants, to examine the psi performance in general. Two additional analyses used only the participants with a high SS-score (SS-score $>$ 2.5) or only those who scored low on the SS-scale (SS-score $<=$ 2.5). These additional analyses were conducted to examine the effect of stimulus seeking on psi performance, identical to the original study [1]. A significance level of α = 0.05 was used for all analyses.

Additionally, an exploratory analysis using mixed-effects models was carried out to study how different variables affect the recall of words. To that end, the package lme4 [15] was used. The estimators that were studied with these models are either our condition variable (whether the word was a target word) or factors that are known to influence the recall of words, such as the SS-score [1], the logarithm of the word frequency [12], the word length centered in the median [16], and word category [16]. Furthermore, by-word and by-subject adjustments were done including the corresponding random-effects structure. A significance level of α = 0.05 was used to determine the significance of the effects.

3. Results

3.1. Replication of Bem's analysis

Three t-tests (investigating the full sample, high SS-score participants, or low SS-score participants) were conducted to examine if target words were more often recalled compared with control words. The results of the current study, as well as the results of the original study, are shown in Table 1. In contrast to the results of the original study, we did not find a significant effect of the target words on the DR-score in the full sample ($M_{DR\%}$ = 0.05, p = 0.97), in participants with a high SS-score ($M_{DR\%}$ = -2.56, p = 0.23), or in the subjects that scored low on the SS-scale ($M_{DR\%}$ = 2.66, p = 0.2). There was also no significant correlation between the DR-score and the SS-score (r = -0.13, p = 0.41).

3.2. Exploratory analysis

As mentioned before, the exploratory analysis is done using linear mixed-effects models.

Random-effects structure
To find the best estimators of the random effects, we tried several models and examined how well they fit the data by looking at the chi-square test. A comparison of some of the models we considered is provided in Table 2. The formula of each of the models that we discuss here can be found in Appendix C. The fixed effects considered in all these models (m1-m7) were the experimental manipulation variable (*WasTarget*) and the intercept (*1*)

From table 2 we conclude that *m3* is the best model. Several combinations of estimators were used, motivated by the hypotheses stated in Section 2.7, but most of them were too complex for the model to even converge. Although *m4* has smaller deviance than *m3* and seems to fit the data slightly better, the Chi-square test indicates that it is not a significant improvement over *m3*.

The random-effects structure of *m3* (Appendix C) is just the interceptor by-subject and by-word: (1|TrialID) + (1|Word).

Experimental condition
The t-test in section 3.1 did not find evidence that the experimental manipulation (being trained on target words) was affecting the probability of recalling those specific words. We can test the same hypothesis now by removing the experimental manipulation variable (*WasRecalled*) from *m3*. We call this model *m3.intercept* (see Appendix C).

If we perform an ANOVA test on *m3* and *m3.intercept*, we observe that *m3* does not present a significant better fit to the data, being *Pr(>Chisq)* = 0.76. The conclusion is the same as in Section 3.1: there is no evidence that target words and control words differed in the frequency with which they were recalled.

Stimulus Seeking Score
Let us now consider Bem's hypothesis that individuals with a higher stimulus seeking score show a stronger precognition effect. If this is the case, we expect to see some

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)	
m1	3	2471.3	2487.9	-1232.6	2465.3				
m2	3	2475.7	2492.4	-1234.8	2469.7	0.0000	0		
m3	4	2442.8	2465.1	-1217.4	2434.8	348.479	1	0.000000003565	***
m4	8	2446.0	2490.4	-1215.0	2430.0	48.775	4	0.3001	
m5	8	2448.5	2493.0	-1216.2	2432.5	0.0000	0	(no convergence)	
m6	8	2450.5	2494.9	-1217.2	2434.5	0.0000	0	(no convergence)	
m7	8	2448.7	2493.2	-1216.4	2432.7	17.347	0	(no convergence)	

Table 2: *ANOVA comparison among models that differed in their random-effects structure to determine the best random-effects structure that will be used for further analyses.*

interaction between *SSScore* and *WasTarget*. To this end we created the model *m3.sss*, which includes the interaction between *SSScore* and *WasTarget* in its fixed-effects structure. An ANOVA test between *m3* and *m3.sss* revealed no significant improvement after the addition of the interaction, with *Pr(>Chisq)* = 0.59.

Fixed-effects structure

After answering the research question using a t-test and a mixed-effects linear model, we are interested in knowing whether we can identify variables in our experimental setup that are influencing the recall of certain words. For that, we expanded the fixed-effects structure to analyse to what extent each variable accounts for the total variance of the data.

For this model we have considered *WasTarget* * *SSScore*; *Category*; *Age* (normalised); *Gender*; *Frequency* (logarithm); and *WordLength* (normalised). The coefficients of the resulting fixed-effect structure are summarised in Table 4.

The results show that the coefficient for *Categoryclothing* significantly differs from zero ($p = 0.04$). A post hoc comparison revealed that the amount of recalled words is higher for the clothing category ($M = 5.38$, $SD = 2.31$) in comparison to the animals- ($M = 4.03$, $SD = 2.04$, $p = 0.01$), foods- ($M = 3.38$, $SD = 1.39$, $p < 0.001$), and occupations-category ($M = 4.25$, $SD = 1.77$, $p = 0.05$). These differences are visually represented in a boxplot in Figure 2. The details of the post hoc comparison are shown in Table 5.

	deviance	Chisq	Df	Pr(>Chisq)
m3				
m3.cat	2421.7	13.11	2	0.0014 **
m3.cat.fr.len	2417.3	4.42	2	0.11

Table 3: *ANOVA comparison between different variations of m3.*

Based on the observation that *Category* has coefficients that are significantly different from zero, and that the p-values for *Frequency* and *WordLength* almost indicated a significant result (see Table 4), we refine our model search by creating two new models: *m3.cat* which only includes *Category* as additional fixed effect, and *m3.cat.fr.len* that includes *Category*, *Frequency*, and *WordLength*. A comparison between these models with the previous model *m3* is shown in Table 3. The results show that the only variable that contributes significantly to the fit of the model is *Category* ($p = 0.001$).

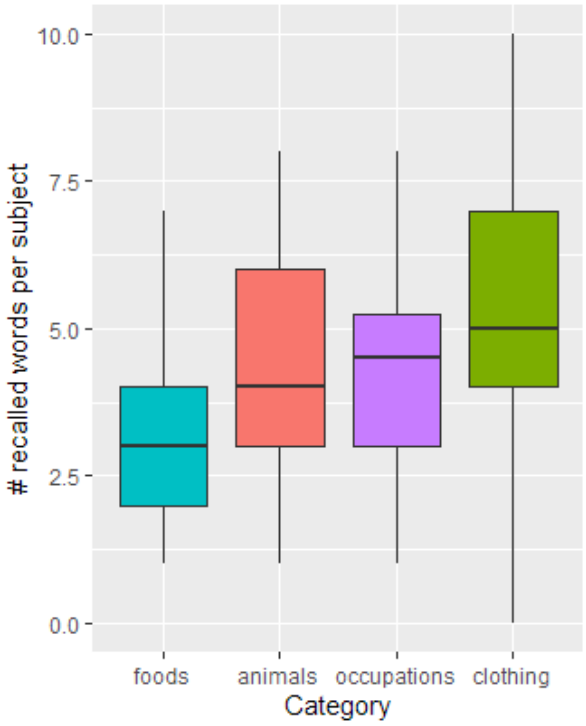

Figure 2: *Boxplot showing the distribution of the frequency of the recalled words per category. The colored boxes represent data points that fall between the 25th and the 75th percentile.*

4. Discussion

The purpose of this study was to replicate an experiment of a previous study [1] that found evidence for the existence of psi. Just like the original study, we investi-

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-211.386	0.76893	-2.749	0.00598	**
WasTargetTRUE	0.36719	0.34753	1.057	0.29071	
SSScore	0.08717	0.11431	0.763	0.44571	
Categoryclothing	0.40726	0.19818	2.055	0.03988	*
Categoryfoods	-0.39249	0.21343	-1.839	0.06592	.
Categoryoccupations	-0.19982	0.23728	-0.842	0.39972	
log(Frequency)	0.13174	0.07723	1.706	0.08803	.
medianLength	0.09159	0.05176	1.769	0.07684	.
Gendermale	-0.08592	0.17354	-0.495	0.62053	
scale(Age)	0.02378	0.08347	0.285	0.77577	
WasTargetTRUE:SSScore	-0.11879	0.11731	-1.013	0.31124	

Table 4: *Fixed-effects coefficients of the full m3 model. Only Category has coefficients that are significantly different from zero (p<0.05).*

	diff	lwr	upr	p adj
clothing-animals	1.35	0.24	2.46	0.010
foods-animals	-0.65	-1.76	0.46	0.425
occ.-animals	0.23	-0.88	1.33	0.952
foods-clothing	-2.00	-3.11	-0.89	<0.001
occ.-clothing	-1.13	-2.23	-0.02	0.045
occ.-foods	0.88	-0.23	1.98	0.174

Table 5: *Post hoc comparison on the recall frequency for the different word categories. In the first column, "occ." stands for "occupations"*

gated precognition by examining the effect of practicing words after a recall task on the amount of recalled practice versus control words. A replication of the analysis of Bem with the results from this study did not find a significant effect of practicing words after a recall task and no correlation between stimulus seeking and psi performance. An exploratory analysis further revealed that only the word category affected whether a word was recalled or not. Words from the clothing category were more often recalled than words from the other three categories (foods, occupations, and animals).

The latter finding, the effect of the word category, is interesting. Our first intuition was to check for an effect of the word length and word frequency as possible sources of confound. However, an ANOVA test between *m3.cat* and *m3.cat.fr.len* revealed that their influence was non-significant. One possible explanation of the observation that clothing items were more often recalled than other words is that participants were most likely wearing at least some of the clothing items that were used as words in this experiment. Thus, instead of just seeing the word, participants could see, and possibly even feel, the item. This is known to improve the recall performance for these words [17]. In contrast, words from the other categories (e.g., a specific food type or animal) were likely less tangible and relatable during the trials.

Overall, the current results contradict the result of the original study of Bem [1]. Whereas Bem found that practice words were significantly more often recalled than control words, we did not find such an effect. This means that, in contrast to Bem's study, our results did not show evidence for the existence of psi. However, just like the results in experiment 9 in the original study, stimulus seeking was not correlated with the psi performance.

The evidence of the current study adds to the growing body of research that indicates the absence of precognition. Although our results are not aligned with those of Bem [1], our study found similar results compared with multiple previous replication studies [2][3][4] that also did not find an effect of practicing words after a recall task. Besides the contradictory results about this topic, the biggest challenge in finding support for precognition, and psi in general, is that it is not in line with our current understanding about time and causality. There is no theory that can explain the existence of the ability to foresee the future. As argued by Wagenmakers and colleagues [18], the evidence in favor of psi must be extremely convincing to "convince a skeptic that the known laws of nature have been bent". Furthermore, the expected consequences of such abilities are generally not present. For example, if people were aware of the future, casinos would go bankrupt very quickly and accidents would be rare.

There could be several different reasons for why we did not find the same results as Bem. First, Bem has been accused of multiple bad research practices (e.g., extensive pilot testing and selective stopping) that could have caused a significant result in his study that will likely not be found in a replication study [10]. Another study pointed out that studies that were not conducted by Bem himself were less likely to find results that support psi [2]. Furthermore, there are some differences between the current and the original study that could have contributed to the inconsistent findings. One example is that we used words that were within the same frequency range, in con-

trast to the words that Bem used. Although this should not affect the psi performance because the target words are drawn randomly, word frequency has been shown to affect performance in free-recall tests [16], thus controlling for it might be a good idea to rule it out of the equation as much as possible.

Another difference and a possible limitation is that our experiment was not conducted in person. This has the consequence that the experimenter can not interfere with the experiment. On one hand, this is an advantage because the view of the experimenter on psi can not influence the results. However, there are also downsides to an online experiment. One downside is that participants might spend more time on the experiment than expected. This was the case for one participant that spent 68 minutes on the experiment, whereas the experiment could easily be completed in 15 minutes. A slightly longer duration than the expected 15 minutes could indicate that participants took their time to complete the tasks in the experiment as good as possible. Nevertheless, a duration of 68 minutes suggests that this person was distracted from the task or took breaks in between. As only one participant spent a remarkably long time on the experiment, it is likely that this did not distort the current results too much.

Another remark about this study is that it is probable that a large percentage of participants were non-native English speakers. This suspicion is based on our own background and the fact that a large part of the participants were our own acquaintances. Since there was no question that asked the participants whether English was their native language, we can not be sure how many people were non-natives and if this had an effect on our results. A previous study [19] suggests that non-native English speakers recall less English words compared with native speakers. This could therefore have been a distorting factor. For future studies, it is advisable to keep track of the native language of participants or only include native speakers.

Finally, a future study could also control more closely for the effect of the word category. The current study found that words from certain categories were recalled more often in comparison with words from different categories. A future replication study could rule the possible influence by word category out by using only one word category.

In conclusion, unlike Bem [1], this replication study did not find an indication that people are influenced by future practice exercises in a recall task. This finding does not support the existence of precognition and psi. It seems that, at least for now, there is no need to question the general beliefs about time, causality, and reality.

5. References

[1] D. J. Bem, "Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect." *Journal of personality and social psychology*, vol. 100, no. 3, p. 407, 2011.

[2] J. Galak, R. A. LeBoeuf, L. D. Nelson, and J. P. Simmons, "Correcting the past: Failures to replicate psi." *Journal of personality and social psychology*, vol. 103, no. 6, p. 933, 2012.

[3] S. J. Ritchie, R. Wiseman, and C. C. French, "Failing the future: Three unsuccessful attempts to replicate bem's 'retroactive facilitation of recall'effect," *PloS one*, vol. 7, no. 3, p. e33423, 2012.

[4] E. Robinson, "Not feeling the future: A failed replication of retroactive facilitation of memory recall." *Journal of the Society for Psychical Research*, vol. 75, no. 904, 2011.

[5] D. Bem, P. Tressoldi, T. Rabeyron, and M. Duggan, "Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events," *F1000Research*, vol. 4, 2015.

[6] C. Honorton, D. B. Ferrari, and G. Hansen, "Meta-analysis of forced-choice precognition experiments (1935–1987)," *The Star Gate Archives: Reports of the United States Government Sponsored Psi Program, 1972-1995. Volume 2: Remote Viewing, 1985-1995*, p. 291, 2018.

[7] J. A. Mossbridge and D. Radin, "Precognition as a form of prospection: A review of the evidence." *Psychology of Consciousness: Theory, Research, and Practice*, vol. 5, no. 1, p. 78, 2018.

[8] M. W. Wagner and M. Monnet, "Attitudes of college professors toward extra-sensory perception," *Zetetic Scholar*, vol. 5, no. 7, p. 16, 1979.

[9] E. P. LeBel and K. R. Peters, "Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice," *Review of General Psychology*, vol. 15, no. 4, pp. 371–379, 2011.

[10] U. Schimmack, "Why the journal of personality and social psychology should retract article doi: 10.1037/a0021524 "feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect" by daryl j. bem," 2018. [Online]. Available: https://replicationindex.com/2018/01/05/bem-retraction

[11] J. Zehr and F. Schwarz, "Penncontroller for internet based experiments (ibex)," *URL https://doi.org/10.17605/OSF.IO/MD832*, 2018.

[12] M. Brysbaert, P. Mandera, and E. Keuleers, "The word frequency effect in word processing: An updated review," *Current Directions in Psychological Science*, vol. 27, no. 1, pp. 45–50, 2018.

[13] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2020.

[14] M. H. Lamers and M. Lanen, "Changing between virtual reality and real-world adversely affects memory recall accuracy," *Frontiers in Virtual Reality*, vol. 2, p. 16, 2021.

[15] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[16] C. R. Madan, "Exploring word memorability: How well do different word properties explain item free-recall probability?" *Psychonomic Bulletin & Review*, pp. 1–13, 2020.

[17] A. Paivio, R. Philipchalk, and E. J. Rowe, "Free and serial recall of pictures, sounds, and words," *Memory & Cognition*, vol. 3, no. 6, pp. 586–590, 1975.

[18] E.-J. Wagenmakers, R. Wetzels, D. Borsboom, and H. L. Van Der Maas, "Why psychologists must change the way they analyze their data: the case of psi: comment on bem (2011)." 2011.

[19] H. Vander Beken, E. De Bruyne, and M. Brysbaert, "Studying texts in a non-native language: A further investigation of factors involved in the l2 recall cost," *Quarterly Journal of Experimental Psychology*, vol. 73, no. 6, pp. 891–907, 2020.

APPENDIX

A. Word list

word	frequency	category
meat	46401	foods
cheese	38025	foods
bread	34840	foods
rice	33676	foods
corn	26609	foods
salad	23010	foods
soup	21273	foods
lemon	17281	foods
vinegar	9110	foods
spice	8795	foods
cereal	7148	foods
tofu	2959	foods
actor	33234	occupations
nurse	28974	occupations
musician	24829	occupations
designer	19814	occupations
consultant	16798	occupations
politician	13369	occupations
waiter	8916	occupations
cleaner	6653	occupations
dentist	6071	occupations
accountant	4598	occupations
tailor	2092	occupations
trucker	1889	occupations
bird	38421	animals
sheep	14372	animals
lion	12969	animals
snake	12901	animals
bee	8674	animals
frog	6230	animals
penguin	4324	animals
donkey	3687	animals
dolphin	2999	animals
panda	2221	animals
snail	1906	animals
hamster	1116	animals
dress	47994	clothing
hat	34222	clothing
coat	28167	clothing
jeans	16585	clothing
gloves	13029	clothing
skirt	11853	clothing
socks	10442	clothing
shorts	10133	clothing
underwear	8622	clothing
scarf	5499	clothing
bikini	3053	clothing
parka	1152	clothing

Table 6: *Words used in our experiment grouped by category with their respective frequency according to the Corpus of Contemporary American English.*

B. Summary of responses

TrialID	Age	Gender	Target.Score	Control.Score	Precog,Score %	SSScore
1	27	female	3	4	-1,22	4,0
3	56	female	9	4	11,28	2,5
4	59	male	10	8	6,25	2,5
5	21	male	10	8	6,25	3,5
6	21	female	8	12	-13,89	3,0
7	28	male	5	7	-4,17	2,0
8	24	female	11	13	-8,33	3,5
9	25	male	6	6	0,00	2,0
10	25	female	6	6	0,00	2,0
11	27	female	11	9	6,94	3,0
12	32	male	6	6	0,00	1,5
13	27	male	6	7	-2,26	4,0
14	25	female	11	10	3,65	4,0
15	26	male	14	14	0,00	3,5
16	28	female	9	8	2,95	3,5
17	23	female	8	8	0,00	2,5
18	31	female	3	4	-1,22	2,5
20	37	female	5	3	2,78	3,0
21	38	female	12	7	16,49	2,5
22	39	female	9	7	5,56	4,5
23	24	female	10	10	0,00	3,0
24	29	male	8	7	2,60	3,5
25	24	male	4	3	1,22	4,5
26	36	male	10	6	11,11	2,5
27	24	male	9	16	-30,38	3,0
28	22	female	8	9	-2,95	2,0
29	24	male	10	10	0,00	1,5
30	27	female	11	10	3,65	2,5
31	55	male	8	12	-13,89	2,0
33	41	female	12	8	13,89	2,0
34	26	male	5	7	-4,17	3,0
35	28	female	11	10	3,65	3,0
36	32	female	10	3	15,80	2,0
37	25	female	11	12	-3,99	3,5
38	25	female	9	6	7,81	2,0
39	37	female	13	11	8,33	2,0
40	70	female	9	14	-19,97	4,0
41	28	male	8	9	-2,95	4,5
42	32	female	11	15	-18,06	2,0
43	39	female	5	5	0,00	1,5

Table 7: *Summary of the results for each participant. Target.Score is the number of recalled words that were part of the target group. Control.Score corresponds to the number of recalled words that were part of the control group. Precog.Score is computed using Bem's original formula. A positive Precog.Score value indicates that target words are more often recalled compared with control words. SSScore is the stimulus seeking score, which ranges from 1 to 5. Note that three participants (TrialID 2, 19, and 32) are already removed due to their abnormal performance.*

C. Glossary of linear models

```
m1: WasRecalled ~ 1 + WasTarget + (1|Word)

m2: WasRecalled ~ 1 + WasTarget + (1|TrialID)

m3: WasRecalled ~ 1 + WasTarget + (1|TrialID) + (1|Word)

m4: WasRecalled ~ 1 + WasTarget + (1 + medianLength |TrialID) +  (1 + medianLength |Word)

m5: WasRecalled ~ 1 + WasTarget + (1 + WasTarget |TrialID) + (1 + WasTarget |Word)

m6: WasRecalled ~ 1 + WasTarget + (1 + WasTarget |TrialID) +  (1 + WasTarget |Word)

m7: WasRecalled ~ 1 + WasTarget + (1 + scale(SSScore) |TrialID) +
     (1 + scale(SSScore) |Word)

m3.intercept: WasRecalled ~ 1 + (1 | TrialID) + (1 |Word)

m3.sss: WasRecalled ~ 1 + WasTarget * SSScore + (1 |TrialID) +  (1 |Word)

m3.cat: WasRecalled ~ 1 + Category + (1|TrialID) + (1|Word)

m3.cat.fr.len: WasRecalled ~ 1 + Category + medianLength + log(Frequency) +
     (1 |TrialID) + (1 |Word)

m3.full: WasRecalled ~ 1 + WasTarget * SSScore + Category + log(Frequency) +
                        medianLength + Gender + scale(Age) + (1 |TrialID) + (1 |Word)
```