KPU Registry ID Number: 1016 Date submitted: 24th August 2015

Study Registration for the Koestler Parapsychology Unit Study Registry

1. The title or name of the experiment (for listing the experiment in the registry).

Experimenter Effect and Replication in Psi Research: Round II of a Global Initiative

2. The name, affiliation, and email address for the lead experimenter(s) for the study.

Marilyn Schlitz, PhD, Senior Fellow, President Emeritus, Institute of Noetic Sciences
marilyn@noetic.org.
Arnaud Delorme, PhD, Visiting Scholar, Institute of Noetic Sciences
adelorme@noetic.org
Daryl Bem, PhD, Professor Emeritus, Cornell University
d.bem@cornell.edu

3. A short description or abstract of the purpose and design of the experiment.

The proposed study seeks to study the replication problem in parapsychology through the examination of experimenter belief in psi. The meta-study involves an international collaboration of teachers, experimenters, and experimental volunteers, who will make use of a standardized psi protocol developed by Daryl Bem that has been the focus of several recent replication attempts and that allows for a systematic collection of data under well-controlled conditions. In particular, Bem's studies were designed to be simple and transparent, requiring no instrumentation beyond a desktop computer, taking less than thirty minutes per session, and requiring statistical analyses no more complex than a t-test across sessions or participants. Specifically, the replication protocol will test the retroactive priming aspect of experiment 4 of Bem, 2011.

A previous study conducted by the Institute of Noetic Sciences investigated experimenter effects using the retroactive priming protocol (Experiment 4) of Bem's series. This study included 12 different laboratories across 32 experimenters and 512 participants. The hypothesis that was assessed on a participant basis did not show a significant psi effect when all the participants were considered. However, this hypothesis did return an effect in the expected direction when we considered only participants who completed it in English (n=193). Also when the statistical power was increased by using a single trial analysis, the primary hypothesis was significant. The results did not support a

correlation between study outcome and experimenter expectancy. Overall, these results support the feasibility of a multi-laboratory collaboration and show that single trial analysis might be more appropriate and powerful to process these types of data.

In the current study, each of 20 participants per experimenter will be seated in front of the computer. Participants will be primed with either a short psi-pro or psi-skeptic paragraph. The type of prime will be counterbalanced across participants. Both texts will be of approximately the same length. They will then be asked to respond to 5 questions designed to assess their belief in psi. For each of the 32 experimenters, half of the subjects will be primed with a psi-pro text and half will be primed with a psi-skeptic text (See Annex 2). After priming and belief assessment, participants then go through a 3-min relaxation procedure automated by the computer program before beginning the task (see Bem, 2011).

Each experimental session consists of 40 trials. In each trial an image is randomly selected and displayed, followed by a randomly selected incongruent or congruent priming word. Participants will be instructed to identify images as "pleasant" or "unpleasant" as quickly as they can; after participants respond, the priming word flashes briefly (at which point they do nothing until the next image appears). A total of 20 "unpleasant" and 20 "pleasant" images followed by a priming word (20 congruent and 20 incongruent) will be shown.

To also test for experimenter beliefs, based on an opportunistic selection of experimenters, our study will prime experimenters in a counterbalanced fashion with a short video (6 minutes, psi-pro or psi-skeptic). Before the video, experimenters will receive a link to a survey and will be questioned about their belief in or receptivity to psi (5 questions). Then they will see the prime video. They will then indicate to us by email that they have seen the video. After waiting 24 hours, we will send them a link to a second survey. In this survey, in order to collect data for a future exploratory language-based analysis and to aid integration of the priming prompts, experimenters will be asked to journal for at least 10 minutes before they are again assessed for belief in psi. They will be asked to journal in response to the following two questions:

> What result do you expect from this study and how can you explain why you think this will happen?
>
> Think of a time when you felt certain of something and describe how you knew or understood it to be certain.

After the journaling, experimenters will again be questioned about their belief in or receptivity to psi (5 questions).

Experimenters will complete these tasks prior to beginning experimental testing. Responses to the psi belief questions (as assessed in the second survey) will be the

independent variable. Experimenters will also receive a prerecorded, web-based experimenter orientation that they will watch in conjunction with the priming video. Participants will be recruited and tested on the retrocausal priming experiment by the experimenter. The results of the psi task – an analysis of the "pleasant" vs. "unpleasant" selection response time – will be the dependent measure for both the psi replication attempts and for the experimenter beliefs.

At the end of the experiment, each experimenter will be asked to fill an exit survey where (again) they will be asked their beliefs about psi, and asked to journal for at least 10 minutes on the same two questions on which they journaled previously. In the exit questionnaire, we will also ask them if they were aware that they had been primed.

4. A statement or list of the specific hypothesis or hypotheses being tested, and whether each hypothesis is confirmatory or exploratory.

The proposed experiment will first be a large scale replication of one of D. Bem's experiments from his landmark article "Feeling the future". In particular, this study will explore one of the most controversial aspects of psi research: precognition. The psi experiment tests the confirmatory hypothesis that memory can "work both ways" by testing whether words can influence the perception of an image —even if the image presentation takes place before the word is given. We will also test the hypothesis that the belief of the experimenter will have an impact on the outcome of a psi task, and the hypothesis that belief of the participant will have an impact on the outcome of a psi task.

This experiment will investigate one confirmatory and three exploratory hypotheses. (1-confirmatory) Replicating the previous study by Bem, response time will be shorter for trials with congruent words than for trials with incongruent words. (2-exploratory) The response time effects (differences) of the participants will be greater if they were with experimenters with positive expectations (belief in psi effects as assessed in the second survey) about the experimental outcome than if they were with experimenters with negative expectations (lack of belief in psi effects). (3-exploratory) The response time effects will be greater for participants with positive beliefs/expectations about psi than for participants with negative beliefs/expectations about psi. (4-exploratory) We will finally test the interaction between experimenters and participants belief in psi with the hypothesis that congruent belief leads to increase image-word congruency effect.

The hypotheses (2, 3 and 4) regarding experimenter and participant belief/expectancy is a planned exploratory analysis.

5. The planned number of participants and the number of trials per participant.

The protocol will involve two levels: (1) Experimenters who will receive a standardized training in the experimental procedure, (2) participants (P) who take part in the psi task.

32 Experimenters will be identified and recruited to participate in the study. Each Experimenter will recruit 20 participants for a study about precognition. This will be a total of 640 subjects who will each perform the psi task once. Each participant will be primed with a psi-pro ($N = 320$) or psi-skeptic ($N = 320$) text, respond to the 5 questions about psi belief (Annex 1) and then be presented with 40 images followed by 40 congruent and incongruent words.

6. A statement that the registration is submitted prior to testing the first participant, or indicating the number of participants tested when the registration (or revision to the registration) was submitted.

We confirm that this registration is submitted prior to the testing of the first participant. The total number of participants will be 640.

To prevent criticism against potential falsification of data on our part, all experimenters will be asked to send the data by email simultaneously to us as well as to a third neutral party not involved in the experiment (Caroline Watt at caroline.watt@ed.ac.uk).

If a dataset is incomplete (less than 20 subjects), we will ask experimenters to complete it. If the experimenter fails to acquire the missing data, all data for the experimenter will be ignored and the experimenter will be replaced by another experimenter. If the experimenter acquires data on more than 20 individuals, only the first 20 will be considered in the analysis.

7. The specific statistical test method that is planned for each hypothesis, including any adjustment for multiple analyses.

Hypothesis 1 - response time will be shorter for trials with congruent words than for trials with incongruent words

Because response-time data are positively skewed, each response time (RT) will be usually transformed prior to analysis using both an inverse transformation (1/RT) and a log transformation (log RT). Trials yielding very short or very long response times will be considered to be spurious outliers and will be excluded from the analysis. Ratcliff (1993) suggested using more than one cutoff criterion to ensure "that an effect is significant over some range of non-extreme cutoffs" (p. 519). We will run four sub-analyses, using both data transformations and two different cutoff criteria for long response times, 1,500 ms and 2,500 ms. In all cases we will use both parametric and bootstrap statistics to assess significance.

The bootstrap will be our primary method as it is more robust than parametric statistics. We will use 2000 bootstrap samples. The bootstrap analysis will consist in computing the null distribution by bootstraping the data appropriately (see also annex 3). If both types of analyses return significant results, this is an additional argument showing the robustness of our results. More specifically, preprocessing will be done as follow:

1. Four statistical tests will be done using two response time data transformation (1/RT and log(RT)) combined with two outlier cutoff criteria (exclude trials with response times >1500 ms and > 2500 ms);
2. Data transformations will be applied to the raw response time for each trial.
3. Trials with errors in judging the image as pleasant or unpleasant will be excluded;
4. Trials with response times <250 ms will be excluded;
5. Participants with judging errors on 25% or more of the trials will be excluded;

In a primary analysis, we will be looking at the average difference in log/inv transformed reaction time between congruent or incongruent trials from all participants run by all experimenters. One-tailed paired t-tests with significance set at p=.05 will be used with the difference between average transformed reaction time for congruent and incongruent trials for a participant as the unit of analysis; The specific p-values from these tests will also be reported. We will run four sub-analyses based on the four data processing methods described above.

In a secondary analysis, we will be pooling all single trials from all participants run by all experimenters. We will consider all the data trials collected as if originating from one single large experiment. For each trial, we will subtract the average reaction time of each subject to avoid bias if some participants have consistently faster reaction times than others[1]. Because it is not possible to perform paired statistics, we will perform unpaired statistics on the collection of single trial response time to assess if there is a difference between congruent and incongruent trials. We will run four sub-analyses based on the four data processing methods described above. One-tailed t-tests with significance set at p=.05 will be used with the transformed reaction time for one trial as the unit of analysis; The specific p-values from these tests will also be reported.

In a third analysis, to provide an analysis that avoids distribution assumptions, we will also compute the percentage of participants who had faster average reaction times for congruent trials than for incongruent trials, evaluated by an exact binomial test with MCE of 50%. One-tailed tests with significance set at p=.05 will be used. The specific p-value from these tests will also be reported.

[1] A similar approach would be to perform an ANOVA and use the 640 subjects as factors in the ANOVA. We will be running that analysis as well.

Hypothesis 2 - The response time effects will be greater for experimenters with positive expectations about the experimental outcome.

Data preprocessing will be performed as for hypothesis 1. For these exploratory analyses, the specific p-values will be reported without pre-specifying criteria for significance or acceptable evidence.

Belief in psi will be assessed by taking the sum of the responses for questions 3 and 4 (the sum will be from 0 to 8). There will be three groups of subjects. The positive group will be those who score in the upper 33% on the 0 to 8 scale (33rd percentile on the distribution of scores that are actually obtained in the experiment) .The negative group will be those who score in the lower 33% on the 0 to 8 scale .The middle group will be everyone else.

Our primary analysis will be looking at the average difference in log/inv transformed reaction time between congruent or incongruent trials from all participants run by all experimenters. We will be running a 1x3 ANOVA with the average difference for each participant as the unit of analysis. The specific p-values from these tests will be reported.

For the single trial analysis (secondary analysis), we will be pooling all single trials from all participants run by all experimenters as if originating from one single large experiment. We will thus be running a 3x2 ANOVA analysis with transformed reaction time for one trial as the unit of analysis. The specific p-values from the ANOVA will be reported.

Hypothesis 3 - The response time effects will be greater for participants with positive expectations about the experimental outcome.

This is the same analysis as for hypothesis 2 except that the beliefs of the participants are now being considered instead of the belief of the experimenters.

Hypothesis 4. Response time effects and interaction between experimenters and participants belief in psi.

We will consider 3 levels in the beliefs in psi for experimenters and 3 levels in the beliefs in psi for participants. For these exploratory analyses, the specific p-values will be reported and criteria for significance are not pre-specified.

For the analysis at the participant level, we will be performing a 3x3 ANOVA analysis (experimenter beliefs x participants belief) on the average difference in reaction time between the congruent and non-congruent stimuli for each participant. The specific p-values will be reported. This will be our secondary analysis.

For the single trial analysis, we will be performing a 3x3x2 ANOVA analysis (experimenter beliefs x participants belief x image congruency) with transformed reaction time for one trial as the unit of analysis. Specific p-values will be reported. This will be our primary analysis.

Finally, we will also separate participants using English primes (if any) from the other participants, and rerun the analysis for each group as a preliminary analysis has shown that language could have an influence on the outcome.

8. The power analysis or other justification for the number of participants and trials.

Bem set 100 as the minimum number of participants/sessions for each of the experiments reported in this article because most effect sizes reported in the psi literature range between 0.2 and 0.3. If d = 0.25 and N = 100, the power to detect an effect significant at .05 by a one-tailed, one-sample t test is .80. Cohen's classic 1988 book (Cohen, 1988) on power analysis has tables of effect sizes for various experimental and statistical designs.

The two priming studies in Bem's 2011 article (Experiments #3 and 4) actually achieved effect sizes of 0.26 and 0.23, respectively. Both experiments had 100 subjects. The current study uses a total of 640 in an attempt to increase the power of the main statistical analysis.

9. The methods for randomization in the experiment.

We will assign a positive or a negative prime to experimenters. For experimenters, the most important for us is to have 16 who are primed positively and 16 who are primed negatively. For every new experimenter, we will toss a coin, heads will indicate a positive prime, and tail will indicate a negative prime. As soon as we reach 16 experimenters who are primed in the same way, the remaining of the experimenters will receive the opposite prime, so that we have 16 experimenters who are primed positively and 16 experimenters who are primed negatively. The professor enrolling experimenters will be blind to that procedure. The assignment of positive and negative prime to participants will alternate for each participant enrolled by an experimenter. Participants and experimenters will be blind to that procedure. Again for each experimenter, we want to balance the number of participants who see a positive prime and the number of participant who see a negative prime.

Computer randomly selects priming words to display after the image. This random selection is performed using the random number generator associated with the presentation software. The presentation software is Real Basic. In this experiment, randomizing was implemented by Marsaglia's PRNG algorithm.

10. A detailed description of the experimental procedure.

The protocol will involve two levels: (1) Experimenters who will receive a standardized training in the experimental procedure, (2) participants (P) who take part in the psi task.

Drawing on existing professional networks of teachers and scientists, 32 Experimenters will be identified and recruited to participate in the study (we are confident we can recruit at least 32 experimenters). They will be selected based on their interest in the study, not by their pre-existing belief in a particular outcome for the replication attempt. Each Experimenter will recruit 20 participants (P) for a study about precognition. This will be a total of 640 subjects. The psi experiment tests the hypothesis that memory can "work both ways" by testing whether words can influence the perception of an image —even if the image presentation takes place before the word is given.

Each person at the two levels of the study (experimenters, participants) will be assessed for their baseline belief in psi phenomena; we will make use of 5 simple questions (see Annex 1) to assess belief in psi. Experimenters will also be asked to assess prior to each participant "How likely do you think it is that this session will produce evidence for precognition?" For each of the experimenters, half of the subjects will be primed with a psi-pro text and half will be primed with a psi-skeptic text (see Annex 2). Experimenters input a participant number into the program before beginning the task—this number determines whether the participant views a psi-pro or psi-skeptic text (e.g. an even number views psi-pro). Both texts will be of approximately the same length. Subjects will then go through a 3-min relaxation procedure run by the computer program (see Bem, 2011) before beginning the task.

Experimenters's will be given a standardized training in the experimental protocol prior to initiating their experiment. A web-based short video will instruct them about the experiment protocol for data collection. Experimenters will download a Windows- or Mac-based experimental program that they will install on a computer to collect data. The program will be identical for all Experimenters. A short survey will be used to assure that the Experimenters understand the procedure. To test for experimenter beliefs, based on an opportunistic selection of experimenters our study will prime experimenters with a short video (6 minutes, psi-pro or psi-skeptic) and then question experimenters about their belief in or receptivity to psi. Experimenters will complete these tasks prior to them beginning experimental testing. To collect data for a later exploratory language-based analysis and to aid integration of the priming prompts, experimenters will also be asked to journal for at least 10 minutes after they have seen the priming video and before they are assessed for belief in psi. They will be asked to journal for at least 10 minutes in response to the following:

What result do you expect from this study and how can you explain why you think this will happen?

Think of a time when you felt certain of something and describe how you knew or understood it to be certain.

Based on a pre-existing study design and replicable experimental series reported by Bem (2011), participants will first be shown images (retroactive priming part of experiment 4 of Bem, 2011). Participants will be told that a word will be flashed on the screen just after they made their judgment of the picture.

Upon entering the laboratory, the participant will be told, "This experiment tests for ESP by administering several tasks involving common everyday words. The experiment is run completely by computer and takes less than 20 minutes. The program will give you specific instructions as you go. At the end of the session, I will explain to you how this procedure tests for ESP."

Participants will be seated in front of the computer. After they have been primed with one of the text paragraphs, the computer program instructs them to respond to the 5 questions shown in Annex 1 to assess belief in or receptivity to psi. Subjects will then go through a 3-min relaxation procedure run by the computer program (see Bem, 2011) and then start the retroactive priming task.

20 positive and 20 negative pictures will be drawn from the IAPS set. 20 positive and 20 negative prime words will be used and a prime will be randomly selected on each trial after the participant has responded to the picture. As a result, congruent trials and incongruent trials are randomly sequenced and do not necessarily occur in equal numbers. This makes it impossible for participants to anticipate the type of trial coming up by knowing the types of trials that have already occurred.

Following completion of the experimental session, the Experimenters will transfer the data files by email to the Institute of Noetic Sciences as well as to a third neutral party. Data files will be archived. At the end of the experiment, each experimenter will be asked to fill an exit survey. When all Experimenters have contributed their data files, they will be analyzed.

Experimenters will not obtain feedback on the performance of the participants they enroll until they have finished collecting data for all participants. This prevents optional stopping in case the data does not fit with their expectation.

References:

Bem, D. (2011) Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology, 100,* 407-425.

Cohen, J (2005) Statistical Power Analysis for the Behavioral Sciences, Psychology press.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. Psychological Bulletin, 114, 510–532.

Annex 1: questions about belief in psi
(numbers next to each response will not be shown to subjects)

1) I often enjoy seeing movies I've seen before

Very untrue
Untrue
Between true and untrue
True
Very true

2) I get bored easily

Very untrue
Untrue
Between true and untrue
True
Very true

Extrasensory Perception (ESP is the Reception of Information Without the Use of the Known Senses. It includes:

- Telepathy: ESP of the Thoughts of Another Person.
- Clairvoyance: ESP of Hidden Objects of Distant Events. (Also called Remote Viewing.)
- Precognition: ESP of Future Events.

3) Do you believe ESP exists?

Definitely does not (0)
Probably does not (1)
Don't know (2)
Probably does (3)
Definitely does (4)

4) Have you had any experiences that you believe were genuine ESP?

Definitely not (0)
Probably not (1)
Maybe (2)
Probably yes (3)
Definitely yes (4)

About yourself
5) Have you ever practiced any form of meditation, self-hypnosis, relaxation exercises, or biofeedback?

No never
Only a few times
Occasionally
Regularly in the past
Regularly now

Annex 2: Priming Texts (participant only sees one or the other)

Comment on psi (ESP) by Michael Shermer, PhD.

(Dr. Michael Shermer is the Founding Publisher of Skeptic magazine, a monthly columnist for Scientific American, a regular contributor to Time.com, and Presidential Fellow at Chapman University.)

"… a meta-analysis of...[psi] experiments and found no evidence for psi, concluding that psi data are non-replicable, a fatal flaw in scientific research. In general, over the course of a century of research on psi, the tighter the controls on the experimental conditions, the weaker the psi effects seem to become, until they disappear entirely. This is a very strong indicator that ESP is not real."

Comment on psi (ESP) by Rupert Sheldrake, PhD.

(Rupert Sheldrake is a biologist and author of more than 80 scientific papers and ten books. He was among the top 100 Global Thought Leaders for 2013, as ranked by the Duttweiler Institute, Zurich, Switzerland's leading think tank.)

"Telepathy, ESP, and psychic/psi phenomena in general are real and backed up by convincing evidence; their investigation deserves to be part of science…I take seriously research within parapsychology. I think there is good evidence for precognitive dreams, and also for presentiment, whereby an emotional arousal can have a physiological arousing effect five or six seconds in advance."

Annex 3: Responses to registration guidelines and to comments during registration review regarding the planned bootstrap methods.

> Guideline: specify which variable is randomly sampled or assigned, and whether the randomization is without replacement (permutation) or with replacement (bootstrap);

Response: We will be using bootstrap in all cases, although in some case, our approach is similar to using permutation. For example, for paired comparisons (2 conditions), we will be bootstrapping the sign of the difference (when computing the t-test – see below), which is equivalent to a random permutation of conditions for each case.

> Guideline: specify the number of samples or simulations that will be done, or the algorithm and criteria for determining the number;

Response: We will be using at least 2000 bootstrap samples. This allows to reach p-value of 0.0005 which we deem as conservative enough for the type of analyses we are running.

> Guideline: specify any constraints on the random sampling of the data, such as sampling within groups for bootstraps;

Response: For comparison between paired groups, we will shuffle the conditions for each trial which is equivalent to bootstrapping the sign of the differences when computing paired t-test value. We do not have experimental comparison involving more than 2 paired groups. For comparison between unpaired groups (2 or more), we will bootstrap this data ensemble where all the data has been pooled.

> Guideline: specify what summary measure will be used as the overall outcome for each replication or simulation (e.g., sum of ratings for all trials, mean value for the participants, mean standardized score, etc.)

Response: The measure being used will be the parametric statistics: paired t-test, ANOVA and repeated measure ANOVA. Except, instead of comparing to the standard parametric curves as when using parametric statistics, the output measure will be compared against the bootstrap distribution.

> Guideline: for bootstrap methods, specify whether a simple bootstrap will be used, or which modification(s) (e.g., parametric, Studentized, bias-corrected, pivotal, etc.);

Response: We will use a simple bootstrap. However, we are considering as exploratory analyses using a winsorized or trimmed version, so that our measure is less sensitive to outliers.

> Comment: For the primary analysis of hypothesis 1, will you randomly sample from the trials and generate the difference for each participant from that, or will you randomly sample from the observed differences for each participant?

Response: We will randomly bootstrap of the difference the sign for each case (or in our case t-test value – the two equivalent). For a more complete treatment of this, see Wilcox http://www.amazon.com/Introduction-Estimation-Hypothesis-Statistical-Modeling/dp/0123869838

> Comment: For the secondary analysis by trials, will the sampling be from the raw data for a participant and the mean of the sample subtracted from these trials, or will the sampling be from the data with the originally observed mean subtracted?

Response: We will be using the second one. Removing the mean from each set of trials could introduce some bias that are hard to predict. We prefer a standard approach.

> Comment: Will trails from all participants be pooled for the samples or will the samples be constrained (balanced) by participant?

Response: Our plan is to bootstrap all trials (congruent and non-congruent) because at this point, we have removed the mean subject reaction time and all trials are on the same footing. It would be also possible to select as many trials from each subject in each bootstrap group. However, this is a more complex procedure, and we believe that people could accuse us of massaging the data.

> Comment: Similarly, will the congruent and incongruent trials be pooled or sampled separately?

Response: For 2 conditions, they will be pooled under the null hypothesis that they originate from the same distribution.