Supplementary Materials

This document contains supplementary material for the paper titled: ‘Raising the value of research studies in psychological science by increasing the credibility of research reports: The Transparent Psi Project.’

[Consensus Design Process](#_m9zbzhex7klz) **1**

[Participants of the Consensus Design Process](#_z6tekbveh7dd) 2

[Stopping Rules of the Consensus Design Process](#_968ww6h2378b) 5

[Supplementary information about the replication study](#_wob1dzcl6h9c) **12**

[Analysis of operational characteristics](#_74sktjaob4jx) 12

[Exploratory Analysis](#_26pt9884xn8e) 21

[Determining the Expected Sample Size](#_1j3l1lyv5cro) 24

[Additional Statistical considerations](#_693rk8yxdrxw) 26

[Additional methodological considerations](#_gsa0y29qa8je) 29

[External Audit](#_8e5a2y1bm5gr) 33

[**Transparency Checklist report**](#_anyy4ueivj9z) **43**

[References](#_7ev8n6a24tp8) **48**

# **Consensus Design Process**

Our aim was to develop a consensus design for a replication study of Bem’s [(2011)](https://www.zotero.org/google-docs/?8Lv8c5) Experiment 1 that is mutually acceptable to all stakeholders of the field, both psi proponents and psi sceptics, containing clear criteria for credibility. To achieve this, we utilized a ‘reactive-Delphi’ process [(McKenna, 1994)](https://www.zotero.org/google-docs/?QxaQRw).

## Participants of the Consensus Design Process

The stakeholders were identified via a systematic review of the literature. The eligibility criteria for panel members are listed below. Both criteria 1. and 2. had to be met for eligibility to be included in the consensus panel.

1. Author of a scientific publication contributing to the debate surrounding Bem and colleagues’ 2011 and 2016 papers [(Bem, 2011; Bem et al., 2016)](https://www.zotero.org/google-docs/?4rlFCn) and closely related topics such as the replicability of psi research, meta-analyses of psi research findings, and methodologies of precognition studies. More specifically

a) authors of scientific papers were eligible, if the paper contains either one of the following:

* + Replication study of one of the studies of Bem [(2011)](https://www.zotero.org/google-docs/?8jmb4O)
	+ Meta-analysis including one of the studies from Bem [(2011)](https://www.zotero.org/google-docs/?tr8aTN) in the analysis
	+ Systematic review of psi literature including Bem (2011) or Bem et al. (2016)
	+ Paper included in the meta-analysis conducted in Bem et al. (2016)
	+ Commentaries or responses to commentaries on Bem (2011) or Bem et al. (2016)
	+ Critical evaluation of the methodologies used in psi studies or the replicability of psi studies in general and citing Bem (2011) or Bem et al. (2016) in the process
	+ Detailed evaluation of meta-analyses of psi studies and citing Bem (2011) or Bem et al. (2016) in the process
	+ Detailed evaluation of the Bem (2011) or Bem et al. (2016) paper, the methods used in any/all of the Bem (2011) studies, or the controversy/discussion it generated on the field.
	+ Detailed evaluation of precognition studies or the controversies they generate in science and citing Bem (2011) or Bem et al. (2016) in the process

b) reviewers of the Bem et al. [(2016)](https://www.zotero.org/google-docs/?MIc3S8) are also eligible, who participated in the public review/commentary process

1. In the past 15 years had at least one peer reviewed publication in a high quality journal rated [Q1](https://researchassessment.fbk.eu/quartile_score) by [Scimago Journal & Country Rank](http://www.scimagojr.com/) in any field. The journal should have Q1 rating in the year that the paper was published. (This publication does not have to fit criteria 1. and does not have to be parapsychology related.)

Participant candidates identified in the systematic review were contacted via e-mail with information about the study and asking to follow a link if they are interested in participating in the Consensus Design Process. The link led to an online form where candidates answered questions regarding their eligibility. Furthermore, the email asked candidates to forward this email to colleagues who they think might be eligible to participate in the panel (eligibility criteria were linked), or to nominate this person by answering to the email. (It was specified in the invitation that participation in the Consensus Design Process does not grant authorship in the resulting publications.)

The online form asked the candidates to give full reference for at least one relevant publication fitting eligibility criteria 1. The publication the candidate provided in his or her answer and other recent publications from the same author were reviewed to determine whether the participant should be classified as psi proponent researcher or a psi sceptic. (Later this categorization was verified by a self-report question, see below.)

We identified 165 researchers during the systematic review, and 4 others were suggested by the contacted candidates (the list of identified candidates can be found in the materials of the study on OSF). No contact information was found or email bounced for 23, so 142 were contacted. Thirty-one candidates completed the eligibility form of whom 29 were eligible to participate. We met our pre-specified goal to have at least 10 panel members represent both psi proponents and sceptics. Prior studies indicate that a sample size of 12-20 or larger is sufficient in most cases to produce a very reliable and stable judgment in consensus studies [(Jorm, 2015; Murphy, 1998)](https://www.zotero.org/google-docs/?qBvy29).

After the conclusion of recruitment, our lab ran a pilot test of the consensus design survey among colleagues and students to get feedback on the procedure before live data collection started.

The first Consensus Design survey package was sent out on the 27th of May, 2017 to all 29 eligible panel members (15 psi proponents and 14 sceptics). The survey package consisted of an online survey and links to the following supporting materials:

* description of the Bem [(2011)](https://www.zotero.org/google-docs/?GsDcAi) original study
* a detailed research protocol of the planned replication study with detailed rationale for any planned deviation from the original design
* background material on the statistical methods intended to be used in the replication study

The online survey asked for:

* rating of perceived level of methodological quality of the protocol (on a 0-9 Likert scale)
* free text response on how to improve study design
* rating of security against questionable research practices (QRPs) (on a 0-9 Likert scale)
* free text response on how to increase secureness against QRPs
* free text response about what is the appropriate level of statistical confidence to seek in the replication study to draw our predefined conclusions

## Stopping Rules of the Consensus Design Process

The following stopping rules were specified before the start of the Consensus Design Process:

1. The study will be considered ‘concluded with unsuccessful recruitment’ if by the end of week 9 of the accrual process there are less than 8 participants representing either the psi sceptics or the psi proponent researchers. The replication study protocol will be finalized without further stakeholder feedback.
2. The study will be considered ‘concluded with consensus’ when consensus is reached about the acceptability of the proposed replication study protocol at any survey stage. We define consensus as a median rating of at least 8 with an interquartile range of at most 2 on both the methodological quality and the secureness against QRP questions of the survey in a survey round in which at least 8 participants submitted valid responses to these two questions among both the psi sceptics and psi proponent researchers. (If the two sides will be largely imbalanced in number of participants, we might consider applying these rules for each group separately, so that the overwhelming support on one side cannot mask the lack of support from the other side.) If this stopping rule is triggered, no new survey rounds will be initiated. The replication study protocol will be finalized based on the revised study protocol that achieved consensus. We will also consider implementing changes suggested in the final round of the survey, but only if this is judged to hold no risk for losing consensus.
3. The study will be considered ‘concluded with no consensus’ if no consensus has been reached at the end of the fourth survey round. If this stopping rule is triggered, no new survey rounds will be initiated. The replication study protocol will be finalized based on the last version of the revised study protocol and the feedback received in the final round of the survey.

We collected responses from 25 stakeholders (12 proponent, 13 sceptics) to the first survey package in 112 days. Once the responses were collected, a result summary has been sent out in order to provide feedback for the panel members. The result summary contains descriptive summary statistics about the numerical ratings. After the first round of the consensus design process the perceived level of methodological quality of the protocol had a median rating of 7 (IQR = 3), whereas median rating of the security against QRPs was 8 (IQR = 2,5).

The proposed design was amended based on the suggestions of the panel members and re-sent for evaluation in the second round of the Consensus Design. The second Consensus Design survey package was sent out on the 27th of September, 2017 to all 25 panel members (12 psi proponents and 13 sceptics) who submitted an answer to the previous round. The survey package consisted of an online survey and links to the following supporting materials:

* results summary of round 1
* all their comments submitted in the previous round and all our replies and actions related to them
* their own ratings given in round 1
* all the anonymised comments and our replies to them
* an executive summary about the updated research protocol
* the revised conclusions

The survey asked for

* rating of perceived level of methodological quality of the protocol (on a 0-9 Likert scale)
* free text response on how to improve study design
* rating of security against questionable research practices (QRPs) (on a 0-9 Likert scale)
* free text response on how to increase secureness against QRPs
* free text response on our pre-specified conclusions
* free text response for nominating one or more research auditors
* rating of personal belief about the likelihood of the existence of extrasensory perception (on a 0-10 Likert scale)

A central feature of the reactive Delphi process is that panel members receive feedback about the responses of others in the previous iteration, and their own previous response. Studies on the role of feedback in consensus methods show that it is beneficial in terms of reaching consensus if not only numerical ratings are fed back to the participants, but also reasons for ratings [(Gowan Jr & McNichols, 1993; Woudenberg, 1991)](https://www.zotero.org/google-docs/?1CI3Rn). Thus, in the Results summary the panel members received the de-identified comments of all panel members in an easy-to-search format, categorized by the topic of the comments and containing our response to the comment. Furthermore, all panel members received personal feedback to their specific comments on the protocol.

On the 24th of November, 2017 we have finished the second round of the Consensus Design Process. The evaluation was completed by 13 sceptics and 10 psi proponents. The stakeholders aggregated ratings were a median of 8 (IQR 1) and 9 (IQR 1); (8.5 [IQR 1] and 9 [IQR 1] for proponents and 8 [IQR 1] and 8 [IQR 1] for sceptics) on the methodological quality and security against QRP rating scales respectively, both reaching the pre-defined criteria for consensus. As the stopping rule for reaching consensus was triggered, no new survey rounds was initiated. The replication study protocol was finalized based on the revised study protocol that achieved consensus. We also implemented changes suggested in the final round of the survey, but only if the suggested changes were judged to hold no risk for losing consensus.

De-identified information of the enrolled panel members can be found in the materials of the study on OSF. Of the final 23 panel member who participated in the second round of the Consensus Design process, the following 22 gave their permission to disclose their membership in the Panel. They are listed in alphabetical order in Table S1.

Table S1

*Panel Members of the Final Consensus Design Process Round Who Gave Permission to Disclose Their Membership*

|  |  |
| --- | --- |
| Name | Affiliation |
| Daryl Bem | Cornell University |
| Dick Bierman | University of Groningen |
| Robbie C.M. van Aert | Tilburg University |
| Denis Cousineau | Université d'Ottawa |
| Michael Duggan | Nottingham Trent University (Ph.D institution) |
| Renaud Evrard | University of Lorraine |
| Christopher French | Anomalistic Psychology Research Unit, Goldsmiths, University of London |
| Nicolas Gauvrit | no affiliation listed |
| Ted Goertzel | Rutgers University at Camden, Emeritus |
| Moritz Heene | no affiliation listed |
| Jim Kennedy | Retired from career involving data analysis for academic, non-profit, government, and industry organizations |
| Daniel Lakens | Eindhoven University of Technology |
| Alexander Ly | University of Amsterdam and Centrum Wiskunde & Informatica, Amsterdam |
| Maxim Milyavsky | Faculty of Business Administration, Ono Academic College, Kiryat Ono, Israel |
| Sam Schwarzkopf | School of Optometry & Vision Science, University of Auckland, New Zealand |
| Björn Sjödén | Göteborg University |
| Anna Stone | University of East London |
| Eugene Subbotsky | Lancaster University |
| Patrizio Tressoldi | Dipartimento di Psicologia Generale - Università di Padova, Italy |
| Marcel van Assen | Departement of Methodology and Statistics, Tilburg University, the Netherlands & Departement of Sociology, Utrecht University, the Netherlands |
| David Vernon | no affiliation listed |
| Eric-Jan Wagenmakers | University of Amsterdam |

#

#

# **Supplementary information about the replication study**

## Analysis of operational characteristics

**Analysis of operational characteristics: Method 1.** To assess the operational characteristics (power and inferential error rates) of our study, we have simulated experiments where data were gathered from populations with a given chance for correct guesses where homogeneous correct guess chance was assumed in the population (i.e. no systematic individual differences in success rate). Matching our sampling plan, we sampled sequentially from this population, and at each predetermined stopping point conducting all four statistical tests included in the primary analysis (the mixed effect logistic regression test and the Bayesian proportion tests with the three different priors). We stopped sampling if any of the stopping rules were triggered. We have simulated multiple scenarios, with the true chance for correct guesses in the population set at 45, 48, 49, 49.5, 49.8, 49.9, 50, 50.1, 50.2, 50.3, 50.4, 50.5, 50.6, 50.7, 50.8, 50.9, 51, 51.1, 51.2, 51.5, 52, 53, and 56 percent. For each of these scenarios, we simulated 5,000 experiments each time drawing a new random sample, and in each of these, the primary statistical inference and the result of the robustness check were recorded. For the scenario where the true correct guess rate was 50 and 51%, we simulated 10,000 experiments, since these thresholds were of particular interest. In this investigation, we found the following operational characteristics for our benchmark effect sizes, 50% correct guesses and 51% correct guesses.

1. When the chance of successful guesses in the population was 50%, the probability of:
	* correctly supporting M0 was 0.962
	* inconclusive study was 0.038
	* falsely supporting M1 was lower than 0.0001 (no falso support for M1 was found in the 10000 simulations)
2. When the chance of successful guesses in the population was 51%, the probability of:
* correctly supporting M1 was 0.999
* inconclusive study was 0.0009
* falsely supporting M0 was 0.0006

(Note that these probabilities are not exact, they are estimated based on simulation.)

This investigation also revealed that our sampling and analysis plan would be able to correctly support M1 with a decent accuracy even if the true correct guess chance was 50.7% in the population (correct inference rate = 88%). However, its sensitivity drops off if the effect is lower than this, and if the true correct guess chance in the population was 50.2% or lower, the study would falsely indicate strong support for M0 in more than half of the experiments. Thus, we need to be aware that extremely small effects might be unnoticed by our study. However, importantly the probability of false support for M1 always remains very low, with p < 0.0002 even at its highest point. The simulations assuming large personal differences indicated a 0.9 probability to correctly support M1 if the true correct guess chance was at least 51%, and the probability of falsely supporting M0 still remained below 0.002 in this case.

The performance of our model was also tested assuming personal differences in guess chance. In the scenario where average population guess chance was set to 51%, for each participant we computer a personal guess chance by adding a random number to this population average drawn from a normal distribution with a mean of 0 and a standard deviation of 0.15. This means that we simulated a population where people have different chance to guess the location of the target correctly, with 10% of the simulated sample being extremely lucky (greater than 70% chance of success), and 10% of the population being extremely unlucky (less than 30% chance of success). The 10,000 simulated studies with this scenario indicated 0.9 probability to correctly support M1, and the probability of falsely supporting M0 being 0.0013.

The results of this analysis of the operational characteristics are visualized in Figure S1. Detailed data on the probability of the different inference decisions for each simulated effect sizes (correct guess chance in the population) can be found in Table 2.



*Figure S1.* The figure depicts the results of the simulation of the operational characteristics of the statistical approach used. Simulated probability of successful guesses in the population is displayed on the x axis, while proportion of statistical inference made is displayed on the y axis.Dark, medium and light gray areas represent the different inferences, support for M0, inconclusive evidence, and support for M1 respectively, based on the simulations.

Table S2

*Operational Characteristics of the statistical approach*

|  |  |  |  |
| --- | --- | --- | --- |
| Correct guess chance in the population | Probability of supporting M1 | Probability of supporting M0 | Probability of inconclusive study |
| 45.0% | 0.0000 | 1.0000 | 0.0000 |
| 48.0% | 0.0000 | 1.0000 | 0.0000 |
| 49.0% | 0.0000 | 1.0000 | 0.0000 |
| 49.5% | 0.0000 | 1.0000 | 0.0000 |
| 49.8% | 0.0000 | 0.9988 | 0.0012 |
| 49.9% | 0.0000 | 0.9932 | 0.0068 |
| 50.0% | 0.0000 | 0.9624 | 0.0376 |
| 50.1% | 0.0010 | 0.8786 | 0.1204 |
| 50.2% | 0.0094 | 0.7004 | 0.2902 |
| 50.3% | 0.0394 | 0.492 | 0.4686 |
| 50.4% | 0.1540 | 0.3024 | 0.5436 |
| 50.5% | 0.3902 | 0.1478 | 0.4620 |
| 50.6% | 0.6534 | 0.0682 | 0.2784 |
| 50.7% | 0.8786 | 0.0256 | 0.0958 |
| 50.8% | 0.9670 | 0.0092 | 0.0238 |
| 50.9% | 0.9938 | 0.0030 | 0.0032 |
| 51.0% | 0.9985 | 0.0009 | 0.0006 |
| 51.1% | 0.9998 | 0.0002 | 0.0000 |
| 51.2% | 1.0000 | 0.0000 | 0.0000 |
| 51.5% | 1.0000 | 0.0000 | 0.0000 |
| 52.0% | 1.0000 | 0.0000 | 0.0000 |
| 53.0% | 1.0000 | 0.0000 | 0.0000 |
| 56.0% | 1.0000 | 0.0000 | 0.0000 |
| Sampled from prior distribution | 0.9855 | 0.0066 | 0.0079 |

The simulation using method 1 also revealed that when the correct guess rate in the population was set to 50% and 51%, the probability for the robustness tests to lead to the same inference as the main analysis was 0.81 and 0.66 respectively (robustness is 0.80 when the correct guess rate was 51.1%). This demonstrates an acceptable power for the robustness analyses as well both if M0 and if M1 was true.

**Analysis of operational characteristics: Method 2.** Operational characteristics were also assessed with another method described by Schönbrodt and Wagenmakers [(2018)](https://www.zotero.org/google-docs/?eHGqJe), in which we use previous information about the size of the effect to inform our simulation about the likelihood of different effect sizes. Following this method, we have run 10,000 simulated experiments using the same analysis and stopping rules as in Method 1, but instead of simulating a fixed effect size for all 10,000 simulations, the true correct guess chance in the population was sampled randomly from a beta distribution with parameters alpha = 829 and beta = 733 in each of these simulations. These parameter values are based on data gathered in Bem’s original experiment 1. Because of the one-sided test, the left-hand side tail of this distribution ends at *p* = 0.5. If we assume that Bem’s original study have drawn an unbiased random sample from the same population that we are going to draw from, this method provides a good estimate of the overall operational characteristics of the study if M1 is true. These operational characteristics are studied parallel with operational characteristics of the simulation assuming M0 is true, in which the correct guess chance in the population is 50%. The inference rates are also included in the last row of Table S2. This analysis indicates that if M1 was true, the study had a 0.986 power, a 0.0066 probability of falsely supporting M0, and a 0.0079 probability of being inconclusive. Operational characteristics if M0 is true are displayed above under the discussion of method 1.

The simulation using method 2 also revealed that when the chance of successful guesses in the population was sampled from the above-mentioned beta distribution, the probability for the robustness tests to lead to the same inference as the main analysis was 0.9686, demonstrating an excellent power for the robustness analyses as well if M1 was true.



*Figure S2a.* The figure displays the distribution of the statistical inferences made about all the simulated samples in which M0 was simulated to be true, that is, the correct probability of successful guesses was set to 0.5. The figure shows that when M0 was simulated to be true, the probability of correctly supporting M0 was 0.963, the probability of inconclusive results was 0.037, and the probability of falsely supporting M1 was 0.0002.



*Figure S2b.* The figure displays the distribution of the statistical inferences made about all the simulated samples in which M1 was simulated to be true using method 2, that is, the correct probability of successful guesses was sampled randomly from a beta distribution based on data gathered in Bem’s original Experiment 1. The figure shows that when M1 was simulated to be true, the probability of correctly supporting M1 was 0.985, the probability of inconclusive results was 0.007, and the probability of falsely supporting M0 was 0.008.

All in all, our simulations reveal a very high power and very low false inference rates, except for extremely low effect sizes.

The analysis of operational characteristics detailed above was run in R version 3.5.3. The computer code of the analysis of operational characteristics is available among the Materials on OSF. The operational characteristics code uses the *progress* (v 1.2.), and *HDInterval* (v 0.2.) packages. The plotting script uses the *HDInterval* (v 0.2.), *reshape2* (v 1.4.3.), and the *tidyverse* (v 1.2.1.) packages.

## Exploratory Analysis

The Consensus Design panel suggested including an exploratory analysis regarding the distribution of the correct guess rate of participants. The goal of this approach is to assess the extent of individual differences in guess rate if there are any. If there are individual differences, it may be the case that not the entire population is capable of better than chance guessing of the target location, but rather just a small subgroup of the population. We will refer to this subgroup as ‘ESP users’ from here on.

In this analysis we assess the stochastic dominance of the empirical distribution of the observed successful guess rate over the expected distribution in a sample taken from a population with 50% successful guess rate. If there is a subgroup of ESP users, we would expect a slightly heavier right-hand side tail for the empirical distribution of the observed successful guess rate than if successful guess rate is homogeneous in the population (see Figure S3).
 First, we will draw 18,000,000 simulated samples (1,000,000 simulated participants) from a population where the success rate is 50% in the erotic trials, and the success rates in this simulated sample will be calculated. Then, we will visualize both this expected and the observed empirical distribution as overlaid histograms to inspect potential differences between the two distributions. Furthermore, we will quantify the difference between the distributions using the Earth Mover's Distance (EMD, also known as the Wasserstein metric).
 Note that in this approach we use the success rate achieved by each participant (just like Bem did in his main analysis). As described in the ‘Treating each trial as independent in data analysis’ section, success counting and premature stopping strategies can be used to inflate success rates within participants. Thus, in this particular analysis, we only use data from participants who finished all 18 erotic trials.
 Premature stopping if a participant has lower than x% success rate right before the final trial can still be used to introduce some bias if we only look at completed study sessions, but it’s impact on the project would be very limited, since this is an exploratory analysis with no influence on the main conclusions. Nevertheless, we will also report the average success rate in sessions that were completed compared to those finished prematurely. This information can help in diagnosing bias from premature stopping related to success rate.



*Figure S3.* This figure displays the expected distribution of successful guess rates observed if M0 is true and the true successful guess rate in the population is 50% (light grey area), and the distribution of successful guess rates observed when M1 was simulated to be true, and the simulated population contained two subgroups, the majority group having 50% successful guess rate, and the minority group (10% of the population) having 60% successful guess rate (dark grey area).

We will report and discuss findings of this exploratory analysis in the results and discussion sections of our paper. However, as this is not a part of our confirmatory analysis and our study is not powered to detect individual differences, our findings in the exploratory analysis will not change our conclusions.

We may conduct other exploratory analyses in the study, however, as mentioned above, they will not replace or be described as confirmatory analyses, and their findings will not influence the conclusions or the abstract of the paper.

## Determining the Expected Sample Size

As mentioned in the statistical analysis plan above, we will use a sequential sampling and analysis plan with predetermined analysis points linked to the total number of completed trials (see above). If all participants finish all 18 erotic trials within their study sessions, we will analyze the data at reaching 1,670, 2,102, 2,534, 2,967, 3,399, 3,831, 4,263, 4,696, 5,128, and 5,560 participants who completed the study. (In case of some participants not completing all 18 erotic trials within their study sessions, these participant numbers will be higher, to reach the predetermined stopping points. See also the Incomplete or missing data section.)

According to the results of the simulation, when M0 was simulated to be true, 75.16% of the experiments stopped right at the first interim analysis point conducted when 30,060 erotic trials were finished, and only 4.91% of the experiments lasted until the maximum sample size was reached (100,080 erotic trials). When M1 was simulated to be true using method 2 mentioned above, 94.75% of the experiments stopped at the first interim analysis point, and only 1.08% of the experiments lasted until the maximum sample size. Method 1 revealed that when the true correct guess chance in the population was simulated to be 51%, 33.55% of the studies stopped at the first interim analysis point, and 1.49% of the experiments continued until the maximum sample size. Information about probability of stopping at different interim analysis points for both method 1 and 2 are listed in Table S3. <https://github.com/kekecsz/Transparent_Psi_Project_scripts/blob/master/sample_size_table.csv>.

Table S3

*Descriptive statistics of the stopping simulations*

|  |  |  |  |
| --- | --- | --- | --- |
| Correct guess chance | Stopped at minimum sample size | Stopped at maximum sample size | 75% of studies stopped at this or earlier sample size |
| 45.00% | 1.0000 | 0.0000 | 37836 |
| 48.00% | 1.0000 | 0.0000 | 37836 |
| 49.00% | 1.0000 | 0.0000 | 37836 |
| 49.50% | 0.9972 | 0.0000 | 37836 |
| 49.80% | 0.9336 | 0.0028 | 37836 |
| 49.90% | 0.8768 | 0.0114 | 37836 |
| 50.00% | 0.7816 | 0.0506 | 37836 |
| 50.10% | 0.6508 | 0.1416 | 62388 |
| 50.20% | 0.4946 | 0.3112 | 136080 |
| 50.30% | 0.3562 | 0.4890 | 136080 |
| 50.40% | 0.2334 | 0.5968 | 136080 |
| 50.50% | 0.1454 | 0.5840 | 136080 |
| 50.60% | 0.1152 | 0.4176 | 136080 |
| 50.70% | 0.1408 | 0.2100 | 111528 |
| 50.80% | 0.2268 | 0.0812 | 86958 |
| 50.90% | 0.3578 | 0.0186 | 86958 |
| 51.00% | 0.4944 | 0.0042 | 62388 |
| 51.10% | 0.6434 | 0.0008 | 62388 |
| 51.20% | 0.7854 | 0.0000 | 37836 |
| 51.50% | 0.9746 | 0.0000 | 37836 |
| 52.00% | 1.0000 | 0.0000 | 37836 |
| 53.00% | 1.0000 | 0.0000 | 37836 |
| 56.00% | 1.0000 | 0.0000 | 37836 |
| Sampled from prior distribution | 0.9585 | 0.0098 | 37836 |

## Additional Statistical considerations

**Rationale for choosing Bayes factors for statistical inference instead of Bayesian parameter estimation.** There are several reasons for choosing the Bayes factor analysis approach instead of the Bayesian parameter estimation as out primary hypothesis testing method. First of all, the consensus panel predominantly was in favor of the Bayes factor approach instead of the parameter estimation approach. This was probably due to the role that the Bayes factor approach played in the statistical criticisms against the findings reported by Bem (2011). This approach lets us use the priors previously proposed in the literature in the commentaries by Wagenmakers and Bem. One of the few cases where the Bayes factor approach is thought to be appropriate is in fact in such research questions, where the null model is meaningful and plausible. Our simulations demonstrate desirable operational characteristics for the Bayes factor approach. Also, our simulation analyses showed that the Kruschke (2018) method requires higher sample size targets to produce the same operational characteristics (inferential error rates) compared to the Bayes factor. This is an important consideration when the execution of the study already requires thousands of participants.

**Rationale for choosing +1% successful guess rate as the smallest effect size of interest.** The smaller the effect we want to be able to detect the larger the sample size required, thus, we needed to draw the line somewhere to ensure feasibility of study execution. Our goal in this study was not to prove or disprove the ESP model. Rather, we wanted to evaluate the likelihood that the results presented in the original study (53% (90% HDI: 51% - 55% )) were biased. The ESP proponents and ESP opponents agreed with using +1% successful guess rate as a minimal effect size of interest in light of the consensus derived conclusions for the study. Specifically, they all agreed that if the study yielded support for M0, the conclusions will include that:

‘…The failure to replicate previous positive findings with this strict methodology indicates that it is likely that the overall positive effect in the literature might be the result of methodological biases rather than ESP. However, the occurrence of ESP effects could depend on some unrecognized moderating variables that were not adequately controlled in this study, or ESP could be very rare or extremely small, and thus undetectable with this study design. Nevertheless, even if ESP would exist, our findings strongly indicate that this particular paradigm, utilized in the way we did, is unlikely to yield evidence for its existence…’

We believe that the +1% smallest effect size of interest (SESOI) threshold proposed in the manuscript is consistent with these conclusion (as shown in the quote, the conclusion allows for the existence of an extremely small effect even if M0 is supported, but also notes that if this would be the case, the current paradigm is very inefficient to provide evidence for it).

**Rationale for choosing the Bayes factor thresholds to be 25 and 1/25.** We have surveyed the consensus design panel about what would they consider as appropriate confidence or decision making criteria in our study, given the conclusions that we aim to draw from our study. After aggregating the responses of the consensus panel, the inference thresholds of BF 25 or 1/25 and p < 0.005 were proposed. These thresholds were deemed to be acceptable by the consensus panel members.

**Rationale for not using Bayesian optional stopping.** There are multiple reasons for using a sequential analysis plan instead of optional stopping. First of all, we use a hybrid frequentist-Bayesian method for statistical inference to increase the robustness of our inference (see reasoning for the use of a mixed model logistic regression together with a Bayesian binomial test in the main text). Optional stopping would very soon decrease the p-value threshold in the mixed-model logistic regression to unattainable values. This could be overcome if we only did the mixed-model analysis at reaching the desired BF threshold. However, this has multiple disadvantages: our simulations show that this way the inferential error rates would be too high in a scenario where there are systematic personal differences in correct guess rate. Furthermore it is not clear how should we proceed if the BF threshold is reached but the *p*-value threshold is not. Second, the benefit of Bayesian optional stopping is not as great in our multi-site study, where there is less central control over participant flow.

## Additional methodological considerations

**Randomization.** Bem (2011) uses two types of random number generators in his study series. An algorithmic pseudo-random number generator (pRNG) using multiply-with-carry technique devised by Marsaglia [(1997)](https://www.zotero.org/google-docs/?ZsM6Wz), and a hardware-based true random number generator (tRNG) called Araneus Alea I [(“Araneus Alea I True Random Number Generator,” n.d.)](https://www.zotero.org/google-docs/?broken=UDt49Z). In Experiment 1 replicated in our study, he used both methods, PRNG to determine the reward for each trial and tRNG to determine the target side. There is no reason given for using both types of randomization. It is stated that using different types of randomization can allow for different mechanisms for the expected effect, but it is not described why is it useful for having both types of RNGs in the experiment, and in the other studies described in the same paper, Bem uses only one or the other technique.

The Alea algorithm was chosen to be used for randomization in our study, which is a modern version of the Marsaglia multiply-with-carry pRNG used by Bem. This is a state of the art pseudorandom number generator (pRNG) passing the BigCrush test [(L’Ecuyer & Simard, 2007)](https://www.zotero.org/google-docs/?broken=EeRdyO). (More information on the algorithm can be found via this link:<https://github.com/nquinlan/better-random-numbers-for-javascript-mirror>.) As described by Bem (2011), the use of a pRNG allows for different mechanistic explanations for a positive effect (such as clairvoyance instead of precognition) because of its deterministic nature. However, our study does not aim to draw a conclusion about the existence of precognition or the mechanisms underlying ESP phenomena. Our goal is to determine whether the higher than chance successful prediction rate of later randomly determined events can be replicated as found by Bem (2011) while using the credibility-enhancing methods described in the paper. Additionally, using tRNG would severely restrict the replicability of our study because it requires specialized hardware which is rarely found at psychology labs.

**Experimenter effects.** Informed consenting, data collection and analysis will be automatized and pre-registered, and the protocol for experimenters is completely manualized. In order to minimize researcher degrees of freedom, the manual has undergone extensive piloting to make sure that it covers every important decision that an experimenter would have to make during the research sessions. Proper delivery of the protocol is assessed through trial session videos submitted by the experimenters. Furthermore, PIs are also separated from the experimental sessions by employing research assistants. These characteristics of the study in combination provide considerable protection against common experimenter effects in experimental research.

However, these interventions do not address the so-called 'psi experimenter effects'. The term psi experimenter effect was introduced by [(Kennedy & Taddonio, 1976)](https://www.zotero.org/google-docs/?3jk3CF) to refer to unintentional psi which affects experimental outcomes in ways that are directly related to the experimenter's needs, wishes, expectancies, or moods. Indeed, if there are ESP effects, one may see several ways in which ESP abilities can unintentionally nudge the results of studies to fit the needs of the experimenter. Because psi effects are supposed to be able to act across both space and time, we see no way to effectively protect our study against these psi-related experimenter influences if they truly exist. In fact, these effects are unmeasurable by nature, because to measure such an effect we would need a study with an experimenter, who him- or herself may influence the outcomes of that measurement with unintentional psi effects. Thus, seeing that we cannot fully control or measure this effect, we are content with only controlling for regular experimenter effects.

At the time of entering the study, all experimenters and site-PIs will complete the Australian Sheep-Goat Scale (ASGS) assessing belief in ESP [(Thalbourne, 2010)](https://www.zotero.org/google-docs/?Va6TuX). Descriptive statistics of the experimenters and site-PIs will be reported in the publication of the study. Also, data on the ASGS total scores of the experimenter overseeing the study session and the site-PI whose laboratory the study session is conducted at will be available for each experimental trial in the data repository containing the raw data of the study. We will not include this variable in the confirmatory or the robustness analyses, but other researchers might find this helpful when conducting exploratory analyses and preparing for future studies.

**Public access and the checker effect.** Some may wonder whether the fact that data will be immediately and publicly accessible may alter the result of the study in any way. There are some reports in the parapsychology literature that indicate that the outcome of the experiment depended on who checked the data first after the experiment [(Weiner & Zingrone, 1986, 1989)](https://www.zotero.org/google-docs/?21EQL4), suggesting that the state of mind of the experimenter (such as expectations) might influence whether a psi effect is found or not. There may be psychological or parapsychological explanation for such an effect as shown above with the psi experimenter effect. One proposed parapsychological explanation for this effect is, that precognition is somehow related to quantum events which are undetermined until an observer actually observes them. This might be problematic in studies where the experimenter is the one who checks the results first. However, in our study, it is the subject who checks his or her precognition result first because they get an immediate feedback of their performance. In fact, this immediate reinforcement is a key feature of the original study. Making this data accessible to others immediately after the participant got performance feedback negates the possibility that the outcome of the trial would be 'collapsed' by a researcher or an outsider.

**Number of trials per participant.** The experimental trials are short and thus it would be possible for a participant to complete hundreds of trials in a relatively short period of time, resulting in a high statistical power. Nevertheless, the overall positive effect in the literature was shown in studies using low number of trials per person. It is possible that increasing the number of trials would result in changes that would mask the effect. For example, it is possible that there would be a habituation to the erotic stimuli if we include more trials, which might decrease the desirability of the incentive, this way decreasing performance. Also, because of the phenomenon, known as ‘psi fatigue’, psi proponents might expect ESP performance to decrease or completely diminish after a while. We have no information on how any of these changes might affect the results in this particular paradigm, which would have made it risky to experiment with such changes. So, we decided to keep the number of trials used in the original study.

**Rationale for not blinding participants or experimenters in the study regarding the nature of the study.** Both participants and experimenters in this study are aware of the hypothesis (although, participants are not made aware specifically that the experiment was designed to test precognition), just like in the original study (Bem, 2011). Participants were told by the experimenters in the original study that this is an experiment about ESP, and their attitude toward ESP was recorded exactly the same way as in our replication. It is unclear how the results would be altered by knowledge of the hypothesis in this study. If there is no ESP, the study outcome cannot be influenced by knowledge about the hypotheses. However, if there is ESP, belief about the possible existence of ESP (expectancy) could play a role in the outcome, or the exposure to the questionnaire could otherwise affect the end result. Both the ESP proponents and skeptics in the consensus panel urged us to keep as close to the original protocol as possible, even if some details seem unimportant for the main purpose of the study. We also see this as useful in averting some of the possible post-hoc criticism related to the study not being “exact replication” if the results turn out to support M0. The experimenters will also have to be aware of the hypotheses, because they will inform the participants about the nature of the study. These complications could be avoided if we fully computerized the briefing just like the data collection process. However, this was rejected by some of the panel members in the consensus panel, because it would take the human contact out of the procedure, which may or may not have an important role in eliciting the effect.

## External Audit

All materials referenced in this section are available in the project’s Audit component on OSF via: <https://osf.io/fku6g/>

**Demonstrating Software, Database, and Server Integrity**

**Pre-study software validation.** The validity of the data collection software has been audited by a programmer (from here on: IT quality expert) who is independent of the programmer of the experimental software. This audit included a visual check of the source code of the data collection software, several test-runs to verify that it works as intended and how it responds to invalid inputs and while being loaded with multiple simultaneous sessions. Special attention has been devoted during the inspection to identify exploitable weaknesses of the software. The IT quality expert also ran a test on the data collection software to verify that the software does not introduce positive or negative bias into the data. The IT quality expert completed a Pre-study software validation report. The issues indicated in this report were addressed by the project’s programmer, after which the IT quality expert re-evaluated the software, and completed the final Pre-study software validation report concluding that ‘I agree with all the solution presented, and there is no evidence of other issues. I think the project is very well structured and itis ready to run.’ The reports are available among the Audit reports on OSF.

**Intra-study software and data integrity.** After it has been released for testing, a copy of the software code has been maintained in a version controlled code repository (GitLab) keeping an audit trail. The code running on the project’s server is synchronized with GitLab. Because of version controlling, this method allows for the verification that the live code running on the server remained unchanged during the project, or if any changes occured, what were they and when did they occur. For security reasons the source code won’t be available publicly until the end of the study, but any possible modification of the code will be traceable after its publication via the repository. The IT quality expert will have access to the GitLab repository containing the source code of the software at all times. We do not anticipate any changes to the code after the pre-study software validation. Nevertheless, the code will be maintained in the repository and if changes are necessary, they will be tracked and thoroughly rationalized in a change log. Change log and the source code will be publically available after the end of the study. Raw data of the study will be continuously uploaded to a publicly accessible data repository (GitHub) during data collection, and a copy of the data will also be kept on the main server of the study.

Records will be kept about the people who had change access to the server after the code of the experimental software was finalized, and why do/did they have access. The system password will be changed when a person who had change access to the server leaves the project. The system password will not be recorded in any form, instead, it will be memorized by the people who have change access to the server. The system password will not be provided to anyone other than the authorized individuals and care will be taken that the password cannot be overseen or overheard by unauthorized individuals. Server access will only be permitted with the approval of the Lead-PI and the project’s programmer. The Lead-PI will keep a log of server accesses authorized by him. Unexpected logins to the server will be noted in the unexpected server event log.

**Checking software, database, and server logs.** The IT quality expert will be responsible to verify software and data integrity throughout the project. The IT quality expert will

1. Check the GitLab repository and the software change log for any indication of inconsistency between the pre-registered, the live reposited, and the live server side versions of the software, and whether these inconsistencies could affect data integrity.
2. Check database integrity for any indication of inconsistencies or changes to the raw data.
3. Check the server-side event log and the unexpected server event log, and assess whether the events (if any) could affect data integrity.

At the end of the study, the IT quality expert will submit a final software and data integrity report consisting of the pre-study software validation report and findings during the checking of software, database, and server logs. This software and data integrity report will be published among the Audit reports on OSF.

The final wording of the IT auditor tasks can be found among the materials on OSF: <https://osf.io/dhvrm>

**Research Audit**

**Selection process.** Research auditors, who are independent of the laboratories directly involved in the replication study, will be selected before data collection starts. We have ask for nominations for potential auditors in the consensus design process. Nominations of individual researchers and audit firms were accepted, self-nomination was also possible. We were asking for nominations of researchers or audit firms who are widely respected for their trustworthiness and who have the necessary methodological knowledge to assess the integrity of the project.

We received 15 nominations for individual researchers. We contacted those 11 who were not members of the Consensus Panel. Of those contacted, three indicated interest for serving as auditors for the study, of whom one eventually withdrew, leaving two candidates, who were judged to be eligible as auditors. Their CVs are shared among the Audit materials on OSF. Parallel to this effort, we also negotiated with an audit firm (Adware Research Ltd.) who submitted a bid for the audit of this project. The first bid was too expensive for the project budget, and the later bid did not include all the services that would have been desired, so we decided to sign a contract with the two individual researchers as auditors instead of contracting Adware Research Ltd.

**Audit tasks.** Auditors will perform an audit of study protocol execution and data integrity at the end of the project. (Auditors will be paid for their services.).

The research auditors will be asked to assess protocol delivery and data integrity by:

1. *Comparing the pre-registered and the live study protocol*, and reviewing differences (if any) and reasoning behind the changes in protocol. The Lead-PI will keep an up-to-date record of the live study protocol currently in use, any differences compared to the pre-registered protocol, and reasoning behind the changes to the pre-registered protocol. These documents will be accessible through OSF. *(Documents: OSF pre-registered protocol, published pre-registered protocol (or registered report), document containing the changes to the pre-registered protocol and reasoning about why they were necessary.)*
2. *Reviewing laboratory logs and reports of protocol deviations (if any).* Experimenters will keep a laboratory log of each research session. Instructions for the contents of the laboratory log are given in the Instructions for experimenters document. Also, Site-PIs will have to report all important deviations from protocol immediately accompanied by a detailed description of what happened to the Lead-PI. The Lead-PI will summarize important protocol deviations in the Protocol deviation report. *(Documents: Laboratory logs, Protocol deviation reports.)*
3. *Checking the software and data integrity report.* The integrity of the data collection software, the database and the server will be validated by an IT quality expert. For more information, see the Demonstrating software, database, and server integrity section above.(*Documents: Software and data integrity report.)*
4. *Reviewing video recordings of the trial study sessions.* Each collaborating laboratory will submit video recordings demonstrating that each of their experimenters is adequately trained and capable of delivering the study protocol as intended. For details, see the Video recording of a mock trial session section in the main text, and the Materials on OSF. The auditors will be asked to review these videos to verify that the laboratories and the experimenters were capable of intended protocol delivery. Given the potentially large number of collaborating laboratories and experimenters, this can be done on a random sample of the videos. (*Materials: Trial session video recordings.)*
5. *Optional additional reviews.* If they want to, the auditors will be free to check any other study documents and materials, and include their findings on them in their final report. *(for example raw data, software source code, documents such as ethics documents, grant documents, reports to the grant organization, all publications resulting from the project, software change log, server unexpected event log, study session registry, etc.)*

Auditors will complete a final report, which will include their findings during each of the above-mentioned tasks. The audit report will be published among the Audit reports on OSF.

The final wording of the IT auditor tasks can be found among the materials on OSF: <https://osf.io/nj68r>

# **Cost analysis of the methodological tools applied in the project**

Table S4. Cost analysis of the methodological tools - single-site project

|  |  |  |  |
| --- | --- | --- | --- |
| Research tool | assistant/data collecting experimenter time | researcher time | programmer time |
| Consensus Design | 160h (rough estimate) for a variety of tasks, including a review to establish potential participant pool, coordination and communication with panel members, creation and distribution of materials, collation, categorization, and distribution of responses from panel members, administration, etc. Tasks detailed in ECO paper. | 8h per panel member (for a small project we can calculate with 8 panel members, 64h), 160h (rough estimate) for a variety of tasks for the coordinating researcher, including IRB proposal, establishing eligibility criteria and search parameters for panel members, designing surveys, digesting responses, responding to responses, and re-designing the research protocol based on the feedback from the panel, administration, etc. Tasks detailed in ECO paper. | - |
| Direct Data Deposition | 2h for testing | 1h for write-up in data management plan and ethics proposal | 4h for implementation in the software |
| Born open data | 3h for creating data dictionary | 3h creating and structuring repository, 1h for incorporating into data management plan, 1h for considering ethical issues and incorporating into ethics proposal | 8h, mainly to handle privacy protection |
| Real-time research report | 2h for testing | 8h for figuring out and writing up content of the report | (32h to understand original code and analysis) 8h to implement the shiny or github app |
| Laboratory logs | 1 hour for testing and feedback, +2 minutes per experimental session for experimenters to enter the log (calculating with 100 sessions, 7h) | 3 hours for creating and revisiong cheklist, 4 hours for checking and reacting to lab notes | 10h for implementing automated data recording, depending on what needs to be recorded |
| Manual for experimenters | 3h per experimenter for comprehensive reading and taking notes, sending feedback (calculating with 2 experimenters, 6h) | 20h for write-up and revisions based on experimenter feedback | - |
| Checklist for experimenters | , +1 minutes per experimental session (calculating with 100 sessions, 1.5h) | 3h for write-up (based on comprehensive manual), and revisions based on experimenter feedback | - |
| Training verified by video recording | 1h per video submitted for the experimenter (calculating with 2 experimenters, 2h) | 6h for write-up and revisions based on experimenter feedback, 0.5h per training video to review and give feedback (calculating with two experimenters submitting two videos each, 2h) | - |
| External research audit | 4h for preparing and providing access to all materials | 4h for figuring out and writing up auditor tasks, 2h for consulting with auditors and assisting them with their inquiries, 16h work per auditor (calculating with 1 research auditor, 16h) | 3h for consulting with the auditors and assisting them in understanding the system and gaining access to the system, 16h for the IT auditor's work |
| Preregistration | 4h formatting and uploading materials | 24h writing up research protocol, 24h write up analysis plan and code | - |
| Open materials | 4h for preparing, uploading, and providing access to all materials | 8h preparing the content of materials in a way that is understandable without contacting the authors | 6h preparing software documentation |
| Tamper-evident software | - | - | 1h setting up unix system logs and gitlab syncronization |

Table S5. Cost analysis of the methodological tools - multi-site project

|  |  |  |  |
| --- | --- | --- | --- |
| Research tool | assistant/data collecting experimenter time | researcher time | programmer time |
| Consensus Design | 320h (rough estimate) for a variety of tasks, including a review to establish potential participant pool, coordination and communication with panel members, creation and distribution of materials, collation, categorization, and distribution of responses from panel members, administration, etc. Tasks detailed in ECO paper. | 8h per panel member (since there were about 30 panel members, this took about 240h in our project), 320h (rough estimate) for a variety of tasks for the coordinating researcher, including IRB proposal, establishing eligibility criteria and search parameters for panel members, designing surveys, digesting responses, responding to responses, and re-designing the research protocol based on the feedback from the panel, administration, etc. Tasks detailed in ECO paper. | - |
| Direct Data Deposition | 2h for testing | 1h for write-up in data management plan and ethics proposal | 4h for implementation in the software |
| Born open data | 3h for creating data dictionary | 3h creating and structuring repository, 1h for incorporating into data management plan, 1h for considering ethical issues and incorporating into ethics proposal | 8h, mainly to handle privacy protection |
| Real-time research report | 2h for testing | 8h for figuring out and writing up content of the report | 8h to implement the Shiny or GitHub app |
| Laboratory logs | 1 hour for testing and feedback, +2 minutes per experimental session for experimenters to enter the log (since there were about 1000 sessions in total, about 33h in total in the TPP) | 3 hours for creating and revisiong cheklist, 4 hours for checking and reacting to lab notes | 10h for implementing automated data recording, depending on what needs to be recorded |
| Manual for experimenters | 3h per experimenter for comprehensive reading and taking notes, sending feedback (since the TPP had about 30 experimenters, 90h) | 20h for write-up and revisions based on experimenter feedback | - |
| Checklist for experimenters | 1 minutes per experimental session (Since there were about 1000 sessions in total, about 16.5h in total in the TPP) | 3h for write-up (based on comprehensive manual), and revisions based on experimenter feedback | - |
| Training verified by video recording | 1h per video submitted for the experimenter (in our project there were about 40 training videos submitted, so this took about 40 hours), 5h for the experimenter coordinator | 6h for write-up and revisions based on experimenter feedback, 0.5h per training video to review and give feedback (in our project there were about 40 training videos submitted, so this took about 20 hours) | - |
| External research audit | 4h for preparing and providing access to all materials | 4h for figuring out and writing up auditor tasks, 4h for consulting with auditors and assisting them with their inquiries, 16h work per auditor (in the TPP there were 2 research auditors so this took 32h in our project) | 3h for consulting with the auditors and assisting them in understanding the system and gaining access to the system, 8h for the IT auditor's work |
| Preregistration | 4h formatting and uploading materials | 24h writing up research protocol, 24h write up analysis plan and code | - |
| Open materials | 4h for preparing, uploading, and providing access to all materials | 8h preparing the content of materials in a way that is understandable without contacting the authors | 6h preparing software documentation |
| Tamper-evident software | - | - | 1h setting up unix system logs and gitlab syncronization |

# **Transparency Checklist report**

Raising the value of research studies in psychological science by increasing the credibility of research reports: The Transparent Psi Project.

Transparency Report 1.0 (full, 36 items)

Zoltan Kekecs; Bence Palfi; Barnabas Szaszi; Peter Szecsi; Mark Zrubka; Marton Kovacs; Bence E. Bakos; Denis Cousineau; Patrizio Tressoldi; Kathleen Schmidt; Massimo Grassi; Dana Arnold; Thomas Rhys Evans; Yuki Yamada; Jeremy K. Miller; Huanxu Liu; Fumiya Yonemitsu; Dmitrii Dubrov; Jan Philipp Röer; Marvin Becker; Roxane Schnepper; Atsunori Ariga; Patrícia Arriaga; Raquel Oliviera; Nele Põldver; Kairi Kreegipuu; Braeden Hall; Sera Wiechert; Bruno Verschuere; Kyra Girán; Balazs Aczel

23 August, 2022

Corresponding author’s email address: [@kekecs.zoltan@ppk.elte.hu](@kekecs.zoltan@ppk.elte.hu)

Link to Project Repository:<https://osf.io/3e9rg/>

**PREREGISTRATION SECTION**

(1) Prior to analyzing the complete data set, a time-stamped preregistration was posted in an independent, third-party registry for the data analysis plan. **Yes**

(2) The manuscript includes a URL to all preregistrations that concern the present study. **Yes**

(3) The study was preregistered… **before any data were collected**

**The preregistration fully describes…**

(4) all inclusion and exclusion criteria for participation (e.g., English speakers who achieved a certain cutoff score in a language test). **Yes**

(5) all procedures for assigning participants to conditions. **Yes**

(6) all procedures for randomizing stimulus materials. **Yes**

(7) any procedures for ensuring that participants, experimenters, and data-analysts were kept naive (blinded) to potentially biasing information. **NA**

(8) a rationale for the sample size used (e.g., an a priori power analysis). **Yes**

(9) the measures of interest (e.g., friendliness). **Yes**

(10) all operationalizations for the measures of interest (e.g., a questionnaire measuring friendliness). **Yes**

(11) the data preprocessing plans (e.g., transformed, cleaned, normalized, smoothed). **Yes**

(12) how missing data (e.g., dropouts) were planned to be handled. **Yes**

(13) the intended statistical analysis for each research question (this may require, for example, information about the sidedness of the tests, inference criteria, corrections for multiple testing, model selection criteria, prior distributions etc.). **Yes**

**Comments about your Preregistration**

No comments.

**METHODS SECTION**

**The manuscript fully describes…**

(14) the rationale for the sample size used (e.g., an a priori power analysis). **Yes**

(15) how participants were recruited. **Yes**

(16) how participants were selected (e.g., eligibility criteria). **Yes**

(17) what compensation was offered for participation. **Yes**

(18) how participant dropout was handled (e.g., replaced, omitted, etc). **Yes**

(19) how participants were assigned to conditions. **Yes**

(20) how stimulus materials were randomized. **Yes**

(21) whether (and, if so, how) participants, experimenters, and data-analysts were kept naive to potentially biasing information. **NA**

(22) the study design, procedures, and materials to allow independent replication. **Yes**

(23) the measures of interest (e.g., friendliness). **Yes**

(24) all operationalizations for the measures of interest (e.g., a questionnaire measuring friendliness). **Yes**

(25) any changes to the preregistration (such as changes in eligibility criteria, group membership cutoffs, or experimental procedures)? **Yes**

**Comments about your Methods section**

No comments.

**RESULTS AND DISCUSSION SECTION**

**The manuscript…**

(26) distinguishes explicitly between “confirmatory” (i.e., prespecified) and “exploratory” (i.e., not prespecified) analyses. **Yes**

(27) describes how violations of statistical assumptions were handled. **NA**

(28) justifies all statistical choices (e.g., including or excluding covariates; applying or not applying transformations; use of multi-level models vs. ANOVA). **Yes**

(29) reports the sample size for each cell of the design. **Yes**

(30) reports how incomplete or missing data were handled. **Yes**

(31) presents protocols for data preprocessing (e.g., cleaning, discarding of cases and items, normalizing, smoothing, artifact correction). **Yes**

**Comments about your Results and Discussion**

No comments.

**DATA, CODE, AND MATERIALS AVAILABILITY SECTION**

**The following have been made publicly available…**

(32) the (processed) data, on which the analyses of the manuscript were based. **Yes**

(33) all code and software (that is not copyright protected). **Yes**

(34) all instructions, stimuli, and test materials (that are not copyright protected). **Yes**

(35) Are the data properly archived (i.e., would a graduate student with relevant background knowledge be able to identify each variable and reproduce the analysis)? **Yes**

(36) The manuscript includes a statement concerning the availability and location of all research items, including data, materials, and code relevant to the study. **Yes**

**Comments about your Data, Code, and Materials**

No comments.

Transparency Checklist based on (Aczel et al., 2020)

#

# **References**

Aczel, B., Szaszi, B., Sarafoglou, A. Kekecs, Z., Kucharský, Š., Benjamin, D., . . . & Wagenmakers, E.-J. (2019). A consensus-based transparency checklist. Nature Human Behavior, 1–3. doi:10.1038/s41562-019-0772-6

[Araneus Alea I True Random Number Generator. (n.d.). Retrieved December 15, 2018, from https://www.araneus.fi/products/alea1/en/](https://www.zotero.org/google-docs/?fZjkuR)

[Bem, D. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524](https://www.zotero.org/google-docs/?fZjkuR)

[Bem, D., Tressoldi, P., Rabeyron, T., & Duggan, M. (2016). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research*, *4*. https://doi.org/10.12688/f1000research.7177.2](https://www.zotero.org/google-docs/?fZjkuR)

[Bem, D., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, *101*(4), 716–719. https://doi.org/10.1037/a0024777](https://www.zotero.org/google-docs/?fZjkuR)

[Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6. https://doi.org/10.1038/s41562-017-0189-z](https://www.zotero.org/google-docs/?fZjkuR)

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Converting among effect sizes. In M. Borenstein, L. V. Hedges, J. P. T. Higgins, & H. R. Rothstein, *Introduction to meta-analysis* (pp. 45–49). John Wiley & Sons, Ltd.

[Bradley, M. M., & Lang, P. J. (2007). The International Affective Picture System (IAPS) in the study of emotion and attention. In J. A. Coan, J. J. B. A. Ph.D, & J. J. B. Allen, *Handbook of Emotion Elicitation and Assessment* (pp. 29–46). Oxford University Press, USA.](https://www.zotero.org/google-docs/?fZjkuR)

[Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89.](https://www.zotero.org/google-docs/?fZjkuR)

[Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193.](https://www.zotero.org/google-docs/?fZjkuR) http://dx.doi.org/10.1037/h0044139

[Gowan Jr, J. A., & McNichols, C. W. (1993). The effects of alternative forms of knowledge representation on decision-making consensus. *International Journal of Man-Machine Studies*, *38*(3), 489–507.](https://www.zotero.org/google-docs/?fZjkuR) https://doi.org/10.1006/imms.1993.1023

[Grünwald, P. (2018). Safe probability. *Journal of Statistical Planning and Inference*, *195*, 47–63.](https://www.zotero.org/google-docs/?fZjkuR) https://doi.org/10.1016/j.jspi.2017.09.014

[Jorm, A. F. (2015). Using the Delphi expert consensus method in mental health research. *Australian & New Zealand Journal of Psychiatry*, *49*(10), 887–897.](https://www.zotero.org/google-docs/?fZjkuR) https://doi.org/10.1177%2F0004867415600891

[Kennedy, J. E., & Taddonio, J. L. (1976). Experimenter effects in parapsychological research. *Journal of Parapsychology*, *40*(1), 1–33.](https://www.zotero.org/google-docs/?fZjkuR)

[Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177.](https://www.zotero.org/google-docs/?fZjkuR) https://doi.org/10.3758/s13423-017-1272-1

[Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual (Rep. No. A-8). *Technical Report A-8*.](https://www.zotero.org/google-docs/?fZjkuR)

[LeBel, E. P., & Peters, K. R. (2011). 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*, *15*(4), 371–379. https://doi.org/10.1037/a0025172](https://www.zotero.org/google-docs/?fZjkuR)

[L’Ecuyer, P., & Simard, R. (2007). TestU01: AC library for empirical testing of random number generators. *ACM Transactions on Mathematical Software (TOMS)*, *33*(4), 22.](https://www.zotero.org/google-docs/?fZjkuR)

[Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2018). Replication Bayes factors from evidence updating. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1092-x](https://www.zotero.org/google-docs/?fZjkuR)

[Marsaglia, G. (1997, September 29). A random number generator for c[Computer software]. Retrieved December 15, 2018, from http://mathforum.org/kb/message.jspa?messageID=1519408&tstart=0](https://www.zotero.org/google-docs/?fZjkuR)

[McKenna, H. P. (1994). The Delphi technique: a worthwhile research approach for nursing? *Journal of Advanced Nursing*, *19*(6), 1221–1225. https://doi.org/10.1111/j.1365-2648.1994.tb01207.x](https://www.zotero.org/google-docs/?fZjkuR)

[Murphy, M. K. (1998). Consensus development methods, and their use in clinical guideline development. *Health Technology Assessment*, *2*(3), 1–88.](https://www.zotero.org/google-docs/?fZjkuR)

[Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4](https://www.zotero.org/google-docs/?fZjkuR)

[Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y](https://www.zotero.org/google-docs/?fZjkuR)

[Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322.](https://www.zotero.org/google-docs/?fZjkuR) http://doi.org/10.1037/met0000061

[Shafer, G., Shen, A., Vereshchagin, N., & Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Statistical Science*, *26*(1), 84–101.](https://www.zotero.org/google-docs/?fZjkuR)

[Thalbourne, M. A. (2010). The Australian Sheep-goat Scale: Development and Empirical Findings. *Australian Journal of Parapsychology*, *10*(1), 5-39.](https://www.zotero.org/google-docs/?fZjkuR)

[van der Pas, S., & Grünwald, P. (2014). Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in nested model selection. *ArXiv Preprint ArXiv:1408.5724*.](https://www.zotero.org/google-docs/?fZjkuR)

[Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638.](https://www.zotero.org/google-docs/?fZjkuR) https://doi.org/10.1177%2F1745691612463078

[Weiner, D. H., & Zingrone, N. L. (1986). The checker effect revisited. *The Journal of Parapsychology*, *50*(2), 85.](https://www.zotero.org/google-docs/?fZjkuR)

[Weiner, D. H., & Zingrone, N. L. (1989). In the Eye of the Beholder. *Journal of Parapsychology*, *53*.](https://www.zotero.org/google-docs/?fZjkuR)

[Wierzba, M., Riegel, M., Pucz, A., Leśniewska, Z., Dragan, W. Ł., Gola, M., … Marchewka, A. (2015). Erotic subset for the Nencki Affective Picture System (NAPS ERO): cross-sexual comparison study. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01336](https://www.zotero.org/google-docs/?fZjkuR)

[Woudenberg, F. (1991). An evaluation of Delphi. *Technological Forecasting and Social Change*, *40*(2), 131–150.](https://www.zotero.org/google-docs/?fZjkuR) https://doi.org/10.1016/0040-1625(91)90002-W