Overview of the preregistration

The current study is party conducted as a thesis project for our bachelor students. Therefore, the preregistration consists of multiple sub-preregistration documents. Specifically, we (AS and SH) wrote the main preregistration, covering the research questions, hypotheses, methods, and analysis plan that we may include in an eventual manuscript.
Additionally, each student involved in the project wrote their own preregistration, that contains of their respective research questions, hypotheses and analysis plan. The main preregistration will follow below; the student preregistration documents can be found in the folder “Student preregistrations”.

Description of the study

In response to the crisis of confidence in psychological science, researchers have initiated various large-scale replication projects (Camerer et al., 2018; Klein et al., 2014, 2018; Open Science Collaboration, 2012). Across these projects, between thirty-six percent and seventy-seven percent of the included studies were replicated successfully. These findings may seem surprising and even disturbing to many. But how surprised should we actually be? Could we perhaps have predicted these replication rates in advance?

Surveys and prediction markets among psychological researchers (i.e., experts) show indeed fairly accurate predictions of the replication outcomes (Dreber et al., 2015; Forsell et al., 2018). However, it remains unclear to what extent predictions of the replication outcome are driven by intuitions about the plausibility of the original finding and their respective test-statistics, and whether accurate predictions of replication outcomes can be made not only by experts in the field (i.e., researchers that have a PhD in psychological research) but also by 'lay people'.

Purpose of the study

The purpose of this study is to investigate how well people form the general public can predict whether a psychological study will replicate or not based on a short summary of the study.

This study consists of a survey, which will take you no longer than 30 minutes.
In the survey, participants are asked to (1) read descriptions of psychological studies, (2) report whether they belief that the study will replicate, and (3) indicate their confidence in their decisions. We will have two conditions. In one condition, the participants will base their decisions on the description of the study alone. In the other condition, the participants will base their decision on the description of the study and the Bayes factor for the hypothesis of interest of the original study.

Research questions

1. Does the inclusion of Bayes factors of the original study plus their commonly used categorical interpretation, improve people's probabilistic predictions of replication outcomes compared to when they are only presented with the study description.
2. Can lay people predict replication outcomes above chance level?
3. Can lay people capture the evidential value of psychological studies, i.e., is there a positive correlation between people's confidence in replicability and the replication effect size?

Expected number of participants

We aim to test a total of 206 participants. That is the estimated number of participants that is required to find strong evidence (BF > 10) in favor for our hypothesis, assuming a moderate

effect size of 0.5, power of 0.8 and a study design that compares two independent groups (i.e., t-test).

Data collection procedure

We aim to recruit 206 participants from (1) the platform Amazon Turk, (2) the online subject pool of first year psychology students, and (3) platforms like Twitter or Facebook.
The participants will be offered either course credits for participation, or a monetary reward.

We will exclude participants if (1) they indicated that they did not understand more than 50% of the descriptions, (2) they have a PhD in psychology (i.e., if they qualify as experts), (3) they do not read the descriptions carefully (i.e., they failed an included attention check), or (4) they are already familiar with the replication outcomes from the studies replicated by Camerer et al (2018) and Klein et al (2018).

Attention check:
We will include one bogus study in our survey. In the description of the study we ask participants to answer 'No' on the question whether the study replicates and indicate a confidence rating of 75%. We consider the attention check as failed, if the participants did not answer 'No', and if their indicated confidence does not fall in the interval between 70 and 80.

Research methodology and design

Note that all study materials, including the reanalyses of the original studies as well as the R code corresponding to our analysis plan is available on our OSF repository (accessible via https://osf.io/x72cy/).

Study design:
In a survey, participants will read short descriptions of psychological studies. In one group, participants will base their decision on a short description of the study. In the other group, participants will base their decision on a short description of the study and the Bayes factor of the hypothesis of interest from the original study. Afterwards, they will indicate whether they believe that the study will replicate, and rate their confidence about their decision.

Dependent variables:
Research question (1): Brier Score for each participant, derived from their predictions across the 27 studies. The Brier score reflects the accuracy of probabilistic predictions, taking into account one's binary predictions (study replicates vs. study does not replicate), as well as the confidence of this prediction (0-100%). As the replication outcome (binary) and confidence are asked as separate questions in our design, the confidence rating will be transformed onto a 0-1 scale, where 0-0.5 reflect the prediction that the study will not replicate, and 0.5-1 reflects the prediction that the study will replicate.

Research question (2): Overall accuracy rate (i.e., probability that participants correctly predict whether any of the studies replicates or not).

Research question (3): Correlation parameter Spearman's rho between the confidence ratings of the participants and the effect size of the study replications.

Experimental conditions:
The experiment will include two experimental conditions; stimulus material will either consist of only a description for each of the original studies, or a description and test statistics (i.e., Bayes factor for the hypothesized effect, plus its verbal categorical interpretation based on Lee and Wagenmakers, 2015, p.105). The participants will be assigned to conditions randomly, such that we have the same number of subjects in each of the two conditions.

Moderators:
We include no moderators in our model.

Control variables:
We include no control variables in our model.

For potential additional exploratory analyses (analysis on subsample) we will ask participants whether they are studying psychology or have studied psychology.

Materials:
Participants that were recruited in the Netherlands could choose whether to take the survey in English or Dutch. The description of the studies in both languages is available in our OSF repository. The study descriptions were written by the authors in English. Then, the descriptions were translated to Dutch, assisted with the translation software DeepL.

Hypotheses of interest

We have three main hypothesis that we will test:

1) Our first hypothesis is that participants who receive a description of the study + the Bayes factor will perform better in predicting the replicability of studies (in terms of their Brier score) than participants who only receive a description of the study.

2) Our second hypothesis states that the accuracy rate (i.e., the probability to predict the replication outcome for all studies correctly) will be higher than chance. Note that we will estimate the accuracy rate separately for each condition, if we find a strong Bayes factor (> 10) in favor for the hypothesis that the groups differ.

3) Our third hypothesis states that lay people are able to capture the evidential value of the study, that is we expect a positive correlation between the (average) confidence ratings for each study and the effect size of the replication. Note that we will compute this correlation separately for each condition, if we find a strong Bayes factor (> 10) in favor for the hypothesis that the groups differ.

Analysis plan

R Code corresponding to our analysis plan is accessible via our OSF repository.

Quality check of data and research design:
Based on the Savage-Dickey density ratio, we will test whether the overall accuracy rate is lower than 50% (Hr) or not (He). If we collect strong evidence in favor for the hypothesis that the overall accuracy rate is lower than chance level, we will conclude that our data has failed our quality check. In this case, we will not perform our confirmatory analyses.

Confirmatory analyses:

Research question 1
We will conduct a Bayesian independent samples t-test on the Brier scores (grouping variable = condition). We assign a default Cauchy prior distribution with a scaling factor of 0.707 on the effect size parameter delta.
Transformation of data:
Note that based on our dummy data analysis we assume that the Brier scores will be right skewed. If this is the case (based on visual inspection of the data), we will conduct a log-transformation of the Brier scores for each participant.

Research question 2
Based on the Savage-Dickey density ratio, we will test whether the overall accuracy rate is higher than 50% (Hr) or exactly 50% (H0). Note that we will estimate the accuracy rate separately for each condition, if we find a strong Bayes factor (> 10) in favor for the hypothesis that the groups differ. To test this analysis we will set up a Bayesian hierarchical model. This model assumes that the accuracy rates per participant are drawn from a group level beta distribution which is characterized by the mean omega and spread kappa. We aim to estimate the parameter omega from the group level beta distribution. We assign a beta(10, 10) prior distribution to the parameter omega (centered around 0.5) and a gamma(0.01, 0.01) distribution on the parameter kappa which is considered as generic vague.

Research question 3
Based on the Savage-Dickey density ratio, we will test whether the correlation between the average confidence ratings for each study and the replication effect size is positive (Hr) or exactly 0 (H0). As we expect a monotonic, but not necessarily linear relation between confidence ratings and effect sizes, we will use the Spearman’s rho correlation. Note that we will compute this correlation separately for each condition, if we find a strong Bayes factor (> 10) in favor for the hypothesis that the groups differ. We will estimate Spearman's rho based on a latent normal inference as proposed by van Doorn et al (in press). We assume a uniform prior distribution for the correlation parameter, that is rho ~ uniform(-1, 1).

Inference criteria:

We will base our statistical conclusions on the Bayes factor. If the data shows strong evidence in favor of our hypothesis (i.e., a Bayes factor of 10 or higher) we consider our hypothesis as confirmed.