Pre Registration

By Chiel van 't Hoff

There is a replication crisis within the scientific community. In a survey 52% of scientists claim there is a significant crisis and about 38% report a slight crisis (Baker, 2016). This is even more strongly the case in psychology where replication efforts by the Open Science Collaboration (2015) only managed to replicate 36 out of 100 studies, the replication effort by Camerer et al. (2018) on studies published in *Nature* and *Science* managed to replicate only 12 out of 21 studies and the Many Labs 2 project by Klein et al. (2018) was only capable of replicating 16 out of 28 studies. There is clearly a crisis if less than half studies of all studies in a field can be replicated. This crisis of reproducibility is leading to a deterioration of trust in psychological research as can be seen by for example this headline from the Independent (2015) "Study reveals that a lot of psychology research really is just 'psycho-babble'". So it can be concluded that the science of psychology is full of studies that cannot be replicated and that changes are needed.

Because of the disappointing results reported by the Open Science Collaboration (2015) more replication efforts were done by Camerer et al., 2018 and Klein et al., 2018). And these replication efforts have indeed assessed whether many studies are replicable and weeded out studies that are not replicable. However as has been noted by Simmons (2011) false positive findings are quite expensive (monetary and time wise), and so is it to correct them, thus it would be quite useful if replicability could be assessed without having to do a replication effort for every single study (possibly only the ones that seem replicable). Thus other efforts are made by for example by Camerer et al. (2018) and Forsell et al. (2018). They made prediction markets and surveys in order to assess prior to replication how much trust experts had in the replicability of a study. The relationship between replicability and trust of experts had quite a high correlation in camerer et al the correlation was 0.54 and in Forsell et al. (2018) 67% of studies were categorised correctly (true positive + true negative). This shows that experts opinions can make good predictions on replicability. Furthermore this can also be used as an additional reproducibility indicator. Furthermore these predictions can be used to assess reproducibility across a research project with multiple replications. Thus further investigation of predictions of replicability would be serviceable.

This study builds further on the study from Camerer et al. (2018) and Forsell et al. (2018). By using similair surveys and a subset of the studies they used. However it differs in three ways from the Camerer et al. (2018) and Forsell et al. (2018) study. First of all this study will assess whether lay people and (mostly psychology) students, instead of experts, can predict replicability. Furthermore, in the study by Camerer et al. (2018) and Forsell et al. (2018) frequentist test statistics were given. However, since frequentist test statistics are probably too difficult to properly interpret for lay people Bayesian test statistics will be used (i.e., the Bayes factor of the original study). The reason that Bayes factors are easier to interpret is because they quantify evidence in degrees (instead of significant not significant) which is more in common with how lay people quantify evidence. Furthermore a personal state of belief is easier to understand for lay people then if a study is replicated an infinite amount of times then it will be significant 5% of the time if there is no effect (frequentist interpretation). Thirdly effect sizes will not be assessed because they are too difficult for lay people and most students to predict. Furthermore this study will have two conditions in order to assess whether adding Bayes factors of the original study improves predictions of replicability.

Current study:

Study type: Survey study
study design: between subjects
research questions:
first research question: Does the inclusion of Bayes factors of the original study increase prediction accuracy?

Hypothesis 2:
- alternative hypothesis (H_1)
the alternative hypothesis predicts that (psychology) students and lay people in the Bayesian statistics group will have a higher prediction accuracy than people in the description group.
-Null hypothesis: (H_0)
The null hypothesis resembles the idealized belief of a sceptic. Thus the null hypothesis predicts the groups will not differ in prediction accuracy.

second research question: Can lay people and psychology students predict replicability?

Hypothesis 1:
- alternative hypothesis (H_1)
The alternative hypothesis predicts that (psychology) students and lay people will predict the replicability above chance level in every group. So correct prediction>0.5.
-Null hypothesis: (H_0)
The null hypothesis resembles the idealized belief of a sceptic. Thus the null hypothesis predicts that (psychology) students and lay people cannot predict the replicability of studies above chance level in any of the groups.

third research question: And do the inclusion of Bayes factors make participants more skeptical?

- alternative hypothesis (H_1)
Under the alternative hypothesis it would be predicted that the participants in the Bayesian group will be more skeptical, because most Bayes factors are relatively low.

-Null hypothesis: (H_0)
The null hypothesis resembles the idealized belief of a sceptic. Thus the null hypothesis predicts the groups will not differ in skepticism.

fourth research question: Will participants in the Bayes factor condition be more likely to say a study replicates if the Bayes factor of the original is larger.

- alternative hypothesis (H_1)
Participants their likelihood of saying a study will replicate and bayes factors of the original study are positively correlated.

-Null hypothesis: (H_0):
Participants their likelihood of saying a study will replicate and bayes factors of the original study are not positively correlated.

Summary and connections of hypothesis:
it is hypothesized that the Bayes factor condition will do better than the description condition. The reason this is hypothesized is because most Bayes factors are very low in the original study. Thus if participants see that a study about which they were not sure whether it would replicate after reading the description they might look at the Bayes factor and than see that it has very weak evidence for an effect which will make them more inclined to be skeptical towards study. And thus participants will be more capable of figuring out which studies will not replicate and thus be more accurate. This also means that it is expected that within the Bayes factor condition there will be a positive correlation between the Bayes factor and the degree to which a participant thinks a study will replicate. Furthermore it should also make people more likely to say that studies will not replicate since a lot of the Bayes factor are incredibly low and provide weak evidence (8) weak evidence) and about 7 give moderate evidence and the rest gives strong or extremely strong evidence.

Manipulated variables: whether the participant is shown bayesian test statistics or not.(condition)

Measured variables.
Our dataset will consist of columns containing the following variables in the following order:
1) participant code.
2) whether the participant is a psychology student.
3-29) confidence rating per study.
30-57) Did the participant think it would or would not replicate. 1 (replicate) or 0 (doesn't replicate).
58-85) Was the participant correct (correct) 0 (incorrect).
86-103) Did the participant understand the description. 1 (yes) or 0 (no)
104) participants will be asked if they were already familiar with any of these studies their replication effort.

Methods:

Participants and sampling plan.
Sample size: at least 60 participants (30 in each condition)
sample size rationale: This sample size was chosen because of the limit in resources, however a minimum of 30 participants per Group is required to have a normal distribution. So 30 per condition was chosen as a minimum.
stopping rule: Collection of data will be stopped at the 22nd of april if the minimum of 60 participants is met. Else participants will be collected for two more weeks or untill the minimum of 60 participants is met. If this doesn't work out within two weeks the collection of participants will be stopped and a lower number of participants will be analysed.

exclusion criteria:
1) A particular study, for a participant, will be removed if the participant filled out that he or she did not understand the description.
2) A study will be removed from the analysis if more than 50% of the participants did not understand it's description.
3) A participant will be removed entirely if he or she does not understand more than 50% of the studies.
4) A Participant will be excluded if he or she already knew the outcome of one of the replications.
5) Phds in psychology

6) If participants failed the attention check (one bogus study within the survey that states that participants have to answer 'No' on the replicability question with a confidence of 75% (also: we allow a small uncertainty here and let values between 70 - 80% pass as answered correctly)

Materials:
the replication survey:
This survey will firstly get ask whether the participant is a psychology student Then the degree of belief in the replicability of 27 studies will be assessed after a short description of the studies. Finally the participant will be asked whether he or she has understood the study description.
Procedure:
Participant will be asked on MTURK to participate in this survey or will be asked to participate for research credits. After that the participants will make this survey on their own computer at home.
Randomization:
Participants will automatically be randomly assigned to a condition when they click on the survey.

Analysis plan

in order to address the First research question. "Does the inclusion Bayes factors of the original study increase prediction accuracy? This can be done with a bayesian t test between the Bayes and description group. The variables of interest are the brier scores for each participant we are assigning a default cauchy prior on the effect size.

in order to address the second research question. "Can lay people and psychology students predict replicability whether the original studies will replicate based on a description of the studies and a Bayes factor?" To conduct this quality check first of all a prior has to be made. The prior will represent that the participants will guess at chance level. The prior will be a beta prior with an alpha and beta of 2 and a dependency of k = 5. After that a parameter estimation will be made of the Bayes factor condition and the description condition.

All of the participants will be assumed to be coins from a single mint in which every participant is a coin which are thrown up 27 times (27 studies) and can land on heads or tails (correct or incorrect). After all of the estimation are done it will be checked whether the median and the 95% credible interval of the posterior distribution is below .5.

in order to address the third research question: does the inclusion of Bayes factors make participants more skeptical?" This question can be answered by doing a Bayesian t test with a default cauchy prior on the effect size coefficient between the Bayes factor and description Group to look at the difference between the amount of times that they thought a study would replicate.

in order to address the fourth research question: Will participants in the Bayes factor condition be more likely to say a study replicates if the Bayes factor of the original study is larger. A Bayesian spearman rho coefficient will be made between the bayes factors and the brier scores. We will estimate spearman's rho using the latent-normal inference as proposed by van Doorn et al. (2017) and it will be assessed with a median and 95% credible interval. We will assign a uniform

prior distribution on the range of -1 to 1 on the parameter rho. It will be expected that the Spearman rho coefficient has a small positive correlation between the brier scores and the Bayes factor.