Preregistration

By Job Bank

Title

Can lay people predict replicability?

Introduction

In response to the crisis of confidence, in psychological science, researchers have initiated various large scale replication projects (see e.g., Camerer et al., 2018; Klein et al., 2018; Open Science Collaboration, 2015). These replication projects were to investigate whether there is indeed a crisis. In the project by the Open Science Collaboration (2015), 36 out of a 100 studies replicated (Open Science Collaboration, 2015). The replication effort by Camerer et al. (2018) managed to replicate 12 out of 21 studies and Klein et al. (2018) successfully replicated 16 out of 28 studies. In other words, in the Klein et al. (2018) and Camerer et al. (2018) research, just above 50 percent of the studies replicated. In the project from the Open Science Collaboration (2015) 36 percent of the studies replicated.

The disappointing replicability rates suggest that researchers need to extensively re-evaluate psychological science. First of all, researchers should conduct direct replications for all findings in psychological science. Romero (2018) provided a solution, for the expensive and time consuming replications, by suggesting the student scheme. The student scheme proposes to make students the main contributors to replication efforts (Romero, 2018).

However, to prevent psychological science from low replicability rates in the future, adjustments in the way of conducting research should be considered. According to Benjamin et al. (2018) the statistical threshold, for a significant p-value, of 0.05 is too weak. They suggest to adjust the threshold to 0.005. Changing the threshold is practically simple, easy understandable for all researchers and might therefore be quickly accepted (Benjamin et al. 2018). Nevertheless, the threshold is still 0.05. The arguments provided for the adjustment of the threshold, provided in Benjamin et al. (2018), are all statistical arguments.

Recent literature by Camerer et al. (2018) and Forsell et al. (2018) showed that experts are pretty accurate in predicting replication outcomes. The experts were provided with information on the hypothesis to be replicated, the p-value of the original result and the sample size of both the original study and the replication (Camerer et al. 2018). In Camerer et al. (2018) the correlation between an experts' trust (in replicability) and the effect size of the replication was 0.54. In Forsell et al. (2018) experts even correctly predicted the replication outcomes for 67 percent of the studies. Furthermore, recent literature found that replication failures are more common in social rather than in cognitive psychology and more in surprising

rather than the more intuitive studies. In this study 'coders' were asked to answer the following prompt: "To what extent is the key effect a surprising or counterintuitive outcome?" Coders read the pre-data collection reports from the study. Responses were provided on a scale from 1 = Not at all surprising to 6 = Extremely surprising (Open Science Collaboration, 2015). This recent literature implies that intuition is of interest in replication outcomes.

The current study aims to answer the question whether lay people can, based on a description from the original study, predict replicability. The subset of replicated studies from the Many Labs 2 project (Klein et al. (2018) and Camerer et al. (2018)) will be used to test this. However, the data from the original studies is reanalysed to compute a Bayes factor. There has been repeatedly argued that the interpretation of Bayes factors is more intuitive than p-values or effect sizes (Wagenmakers et al., 2018). Therefore, the author believes that providing lay people with Bayes factors is more intuitive and therefore more accurate than providing p-values or effect sizes. This study differs from previous research in three ways. First of all, lay people will predict replicability instead of experts. Secondly, participants will be provided with Bayes factors instead of p-values. Thirdly, one of the conditions will only provide a description of the original study and no statistical information.

If the results from this study show that lay people can predict replicability, based on a description from the study, this indicates that replication is linked to intuition. Since predictions are based on a description, which activates intuition. Therefore, this study could implicate that a more counter-intuitive study, fails to replicate more often. If this is the case, it needs stronger evidence is to prove the theory. This is yet another reason, at least for the more counter-intuitive studies, to adjust the threshold for a significant p-value to 0.005.

Current study

Study type: Survey study
Study design: between subjects design

Manipulated variables

Participants will be assigned to one of two conditions. The first condition is the description condition. Participants in this condition will only receive a description of the study. The second condition is the statistics condition. In the second condition participants will receive a description of the study and a Bayes factor.

First research question

Does including a test statistic (i.e., Bayes factor) increase prediction accuracy?

Null hypothesis (H_0)

The null hypothesis predicts including a Bayes factor does not increase the prediction rate.

Alternative hypothesis (H_a)

Lay people who receive a description and a Bayes factor, will have a higher prediction accuracy than lay people who will predict replicability on the description alone.

Second research question

Can lay people predict whether a study will replicate?

Null hypothesis (H_0)

The null hypothesis predicts that lay people predict the replicability of studies at chance level.

Alternative hypothesis (H_a)

The alternative hypothesis predicts that lay people will predict the replicability above chance level. Thus, the success rate for a prediction should be higher than 50 percent. According to the Open Science Collaboration (2015), replication failures are more common in less intuitive studies (Open Science Collaboration, 2015). This would suggest that lay people will be able to predict the replicability above chance.

Measured variables

Our dataset will consist of columns containing the following variables, in the following order:

1) Participant code.

2) Psychology student. Yes (1) or no (0).

3-29) Degree of belief in replicability. Lowest (0) to highest (100).

30-57) Prediction whether the study replicates. Correct (1) or incorrect (0).

58-85) Understands the description of a particular study. Yes (1) or no (0).

86) Already familiar with any of the replication outcomes of the studies. Yes (1) or no (0).

Method

Participants and sampling plan

Sample size

At least 60 participants (30 in each condition).

Sample size rationale

The relatively low sample size was chosen because of limited resources.

Stopping rule

Collection of data will be stopped on the 22nd of April, on condition that at least 60 surveys are collected. Else, data collection will continue until 60 surveys are collected with an ultimate deadline on the 1st of may. Then, collection of data will be stopped and the analysis will be done with the collected responses.

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

First of all, the survey will gather some personal information from the participant. This will include whether the participant is a psychology student and whether he or she is doing/ has done a PHD in psychology. Besides that, the survey includes a description of 27 studies (and a bogus study that serves as attention check), gathered from the replicated studies from Camerer et al. (2018) and Klein et al. (2018). This description is inspired on the study description, given in Camerer et al. (2018) or Klein et al. (2018). The description was adjusted to increase the comprehensibility for lay people. Participants will be asked for their degree of belief in the replicability of a study. This degree of belief is scaled from 1 to a 100. Afterwards, the participant will be asked whether they understood the description of the study. At the end of the survey the participant will be asked whether he or she already knew a result from one of the replications.

Procedure

The survey will be placed on MTURK, an online data gathering website, where participants will be paid for filling out the survey. Furthermore, the survey will be placed on the website from the University of Amsterdam where students can fill out the survey, to receive research credits. Participants will participate on their own (facilitated) computer.

Randomization

Participants will be randomized through qualtrics.

Plan for the analyses

Quality check

Before the analyses are executed, a quality check will be done. If the prediction from lay people is below chance level, doing further analyses is not of interest. This could indicate that the participants got confused by the authors' instructions and/ or the descriptions from the studies. Another cause could be misinterpretation of the Bayes factor.

Statistical method

To conduct the quality check, the author will estimate the parameter theta that lay people can correctly predict the replication outcome of psychological studies. We are assigning a beta prior distribution on theta with the shape parameters alpha and beta, both set to 2. This represents a probability density function which is centered on chance level. Afterwards, the author can estimate the parameter for the overall accuracy (on prediction) across

conditions. If the median and the 95 percent credible interval of the posterior distribution are below 0.5, further analyses are not of interest.

First research question

Research question

Does including a test statistic (i.e., Bayes factor) increase prediction accuracy?

Null hypothesis (H_0)

The null hypothesis predicts that including a Bayes factor, does not increase the prediction rate.

Alternative hypothesis (H_a)

Lay people who receive a description and a Bayes factor, will have a higher prediction accuracy than lay people who will predict replicability on the description alone.

Statistical method

A Brier score is computed for all participants in each group. The Brier score is a proper score function that measures the accuracy of probabilistic predictions. It does not only take success rate into account but also the degree of belief of a participant ("Brier score," n.d.). A Bayesian independent samples t-test, with a default cauchy prior on the effect size, will be executed to compare the mean Brier scores for the two conditions.

Second research question

Research question

Can lay people predict whether a study will replicate?

Null hypothesis (H_0)

The null hypothesis is that lay people predict the replicability of studies on chance level.

Alternative hypothesis (H_a)

The alternative hypothesis is that lay people will predict the replicability above chance level.

Statistical method

A Bayesian hierarchical model will be set up. It could be compared to multiple coins in a single mint (Kruschke, 2014). All participants (coins) go through the 27 studies (the mint) and either predict, each study, right or wrong (land heads or tails). The prior, a beta distribution, is gently peaked around chance level with K = 5 and the shape parameters alpha and beta are both 2.

The hierarchical model is able to provide the accuracy rate for lay people on predicting replicability. The two conditions could either be compared or merged. This depends on

whether the first research question provides strong evidence (Bayes factor > 10), for a difference in condition.

Literature

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2, 6.

Brier score. (n.d.). In Wikipedia. Retrieved March 30, 2019, from https://en.wikipedia.org/wiki/Brier_score

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2, 637.

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., ... & Dreber, A. (2018). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*.

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443-490.

Lilienfeld, S. O., & Waldman, I. D. (Eds.). (2017). Psychological science under scrutiny: Recent challenges and proposed solutions. John Wiley & Sons.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716.

Romero, F. (2018). Who Should Do Replication Labor?. *Advances in Methods and Practices in Psychological Science*, 1, 516-537.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychonomic Bulletin & Review, 25, 35-57.