Preregistration

by Philip Kleinbrahm

Title: Replication crisis revisited – Investigating the link between QRPs and replication rate and the accessibility of statistical figures for a broader audience

Original Experiments

Within the last decade various replication projects have come forward shedding light on a what appears to be a structural problem of replicability within psychological research. This study focuses on two of the main projects which have been reanalysing important and influential studies in our field of research.

Klein et al. (2018) conducted a research called 'Many Labs 2: Investigating Variation in Replicability Across Sample and Setting'. They sampled from renowned psychological studies and tried to reproduce findings from 28 studies. In order to achieve this goal and let the procedure be as reliable as possible, all replications were direct and performed with very large sample sizes given standards for psychological research, in total 15,305 participants from 36 countries contributed to this project. 21 out of the 28 studies had smaller effect sizes than originally reported and 9 showed effect sizes opposite in direction to the original study. Overall, 15 out of the 28 studies provided evidence for the original results and were counted as replication successes. This amounts to a total replication success rate of approximately 54%.

Camerer et al. (2018) investigated 21 studies from the renowned journals `Science` and `Nature`, trying to replicate them. Additionally to this, the researchers compared the findings of these replication efforts with expert opinions (i.e., researchers that have a PhD in psychological research) . Short descriptions of the studies were provided to the experts and interestingly a strong correlation of in between roughly 0.5 to 0.7 between expert's estimations and actual replication results emerged. This shows that many of the experts were able to correctly predict whether studies replicate or not by having a mere look at heavily condensed information with quite a high accuracy.

Current Study

Study type. Survey study + Experimental study

Primary research question. Is it feasible to devise a checklist for QRPs that reliably predicts research validity/strength of evidence?

Hypotheses of interest.

- Main hypothesis (H_m):
  Questionable Research Practices are mainly bad because they strive to inflate findings or purely construct them out of thin air. To do so, one often needs to be sloppy, dogmatic or unaware of conceptual problems, which eventually should leave traces in the published paper. Having crafted a sound QRP checklist, the resultant scores for each research should positively covary with the Bayes factor of the given study.
- Null hypothesis (H_0):
  Assumes there is no correlation/QRP checklist is worthless.

Secondary research question. Can Bayes factors facilitate lay people's understanding of validity of psychological studies?

Secondary hypotheses. Bayes factors have prominently been advocated as being more intuitive than regular frequentist p-values. If so, this notion should extend to the general population.

Experimental design. Two sample, a between-subjects design.

Lay people will be randomly assigned to two conditions, one in which they will be provided a brief study description and one in which this study description will be accompanied by a Bayes Factor of this study. Consecutively they will be asked to provide an estimation whether the study will replicate and how confident they are in this estimation.

Manipulated variables. Presence of Bayes factor in the description

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

- **(1)** Participant code
- **(2 - 28)** Participants binary predictions for replication success for each of the provided 27 studies (yes/no)
- **(29-45)** Participants confidence in their prediction (in percentage)
- **(46-72)** QRP checklist score

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 participant will be removed entirely if he or she does not understand more than 50% of the studies.

3) A Participant will be excluded if he or she already knew the outcome of one of the replications.

4) PhDs in psychology

5) 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

Studies. Studies were drawn from the aforementioned replication projects. We selected studies which were easy to be re-analysed in a Bayesian framework within JASP, spanning major findings from t-tests, to ANOVAs and Chi-square tests.

QRP checklist. Diverting from a Transparency checklist for aspiring collaborators to the open science project, we will construct a scoring checklist for QRPs. As the transparency checklist is designed to guide researchers in the process of conducting a research, it focuses on how to avoid QRPs and consequently forms the basis for Open science and an easy-to-follow understanding of what the research team has done to arrive at the ultimate research paper. Thus, it can be viewed as somewhat a logical inverse of how to detect QRPs in research papers which not have followed this transparency code. Absence of open reports of important statistical figures, complete datasets and possible sources of bias and such forms a good ground to place mistrust on when reading through a research article. Often though, as one can easily imagine, it can be hard to detect absence of information which has just not been provided, forming a conundrum for the critical reader. Additionally to the transparency guidelines that will be incorporated, signs of potential fraud (i.e. p-hacking and such) will be included, to further solidify the validity of our QRP checklist. This is due to the assumption of the transparency checklist that the researcher applying it is trying to be transparent and not fishing for positive results to get published.

Study descriptions. Loosely related to the descriptions from Klein et al. (2018) and Camerer (2018) et al. we summed up the studies we selected to be understood by a broader public than only PhDs in psychological research .

Procedure. Participants will take an online survey, in which they are briefly introduced to the concept of directly replicating studies. For half of them, a short description of Bayes factors also will be included. Then they have to read the descriptions of our 27 studies and give their estimation whether or not the study will replicate and indicate how confident they are in their decision . Descriptions will be available in English and Dutch to guarantee the highest possible understanding for any given individual taking part in this study. Exclusion criteria are set in place to ensure that ultimately data only incorporates serious responses of the outlined lay people population.

Randomization and blinding. Participants will be randomly assigned to either the group of participants which are provided the study description only or the description + Bayes factor condition.

Analysis Plan

Primary Research Question

To address our primary research question a Spearman correlation will be computed to link the score of the QRP checklist to the Bayes factor of the original study. Logarithmic transformation of Bayes factors will be performed to handle the large variation of this measure.

Statistical models. We aim to quantify evidence for either the null or main hypothesis using a Bayes factor test for continuous probabilities.

Prior specification. Since this is the first endeavour to accomplish arriving at a universal QRP checklist, priors are relatively uninformed. The null hypothesis is a single point and thus does not require a prior. The prior for the main hypothesis is a flat prior ranging from zero to 1. We will compute a Savage Dickey density ratio for null hypothesis testing.

Secondary Research Question

In order to test whether lay people's success rate of guessing replication success increases when having been provided a Bayes factor, we firstly will compute overall comprehension scores for participants in the 'description only' condition and in the 'Bayes factor' condition.
To account for interpersonal differences in overall accuracy and overall confidence levels, we will compute an overall comprehension score per person dependent upon average confidence level and the standard deviation hereof per individual. The ultimate comprehension score per participant per item will be computed as follows. If the item was answered with above-average uncertainty, the inverse of the (*deviation measured in standard deviation +1*) will be used as a weight to account for this uncertainty and decrease the relative influence of the item on the overall comprehension score. If the item was answered with below-average uncertainty, (*deviation measured in standard deviation +1*) will be used as a weight to account for this comparably strong belief and increase the relative influence of the item on the overall comprehension score. Correct answers on a given item will be counted as positive, incorrect answers will be deducted from the overall comprehension score. The 2 resultant distributions of comprehension scores for our 2 conditions will be compared utilizing a Bayesian one-sided t-test. Here it will be checked whether the 'Bayes factor condition' produced higher scores than the 'description only' condition.

To further investigate whether there is an interaction between item-difficulty (i.e. comprehension difficulty of a given study) and personal performance, we will utilize a hierarchical model that will capture the dependence of these two parameter spaces. In addition to the measures introduced above, the item-difficulty parameter will be computed by multiplying variants of average accuracy and uncertainty of responses per item. This will give insight into whether the Bayes condition significantly changes behaviour on studies which are on the extreme ends of the scale, either hard to understand or easy. Because this method places special emphasis on the extremely hard to grasp studies, it is not advisable to exclude studies which have not been understood even by a majority. A small increase in understanding of these studies still can hint at an exceptional quality of our manipulation and should be included into analysis. Prior distributions will be uninformed, given the pioneering role of this research.

References

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*(4), 443-490.

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*(9), 637.

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*(1), 35-57.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... & Meerhoff, F. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic bulletin & review*, *25*(1), 58-76.

Appendix A: Bayes factor interpretation guideline

This table provides a rough guideline for the interpretations of Bayes factors. The interpretation guideline was obtained from Lee & Wagenmakers (2014, p.105).

Bayes factor BF_{10}	**Strength of evidence**
> 100	Extreme evidence for H_1
30 - 100	Very strong evidence for H_1
10 - 30	Strong evidence for H_1
3 - 10	Moderate evidence for H_1
1 - 3	Anecdotal evidence for H_1
1	No evidence
⅓ - 1	Anecdotal evidence for H_0
1/10 - 1/3	Moderate evidence for H_0
1/30 - 1/10	Strong evidence for H_0
1/100 - 1/30	Very strong evidence for H_0
< 1/100	Extreme evidence for H_0