Review of *Evaluating the Replicability of Social Priming Studies*

(ReproducibiliTea UvA; @UvARepTea)

In general, we are very positive about the paper. We find the data presented by the authors to be so conclusive that the statistical procedures are almost unnecessary. Nevertheless, we would like to point out some aspects of the paper that could be still improved.

The replication study analyses are based on the first large-scale replication effort in psychology (Open Science Collaboration, 2015) and do not reflect the latest development in the evaluation of replication success. As the authors say, there are many ways to measure replication success, and there is not necessarily one right way, but at least two of the approaches used in the paper have been severely criticized in the literature.

- Testing whether the 95% confidence interval of the replication effect estimate includes the original effect estimate is flawed because it does not take the uncertainty of the original effect estimate into account. A better approach is to compute a prediction interval for the replication effect estimate, which takes into account the uncertainty from both estimates (Patil et al., 2016, Pawel and Held, 2020). This approach has been used in the analysis of economics, social sciences, and cancer biology replication projects.
  Regardless of whether the confidence or the prediction interval approach is used, if the estimate is not contained in the interval this does not mean that the replication was successful but only that one failed to reject that the underlying effect sizes are the same (some authors say that the replication results are “consistent” with the original study). Put simply, small-N original/replication studies would otherwise have the advantage in replicability according to this method because it is likely to have a wider prediction/confidence intervals than large-N studies.

  Patil, P., Peng, R. D., & Leek, J. T. (2016). What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. Perspectives on Psychological Science, 11(4), 539–544. https://doi.org/10.1177/1745691616646366

  Pawel S, Held L (2020) Probabilistic forecasting of replication studies. PLoS ONE 15(4): e0231416. https://doi.org/10.1371/journal.pone.0231416

- The McNemar test (comparing the proportion of significant studies differs in replication and original studies) suffers from a similar issue as the confidence interval method. It does not take the power of either group of tests as well as potential heterogeneity into account. Therefore, even if the same true effect size existed in both sets of studies, the proportion of significant p-values cannot be expected to be the same if the sample sizes in original and replication pairs differed.

  Jacob M. Schauer. Kaitlyn G. Fitzgerald. Sarah Peko-Spicer. Mena C. R. Whalen. Rrita Zejnullahi. Larry V. Hedges. "An evaluation of statistical methods for aggregate patterns of replication failure." *Ann. Appl. Stat.* 15(1), 208 - 229. https://doi.org/10.1214/20-AOAS1387

Based on these two observations, we think that the manuscript may benefit from a more thorough discussion of the statistical power of replication studies.

The authors repeatedly use I^2 as a measure of heterogeneity. I^2 is not an absolute measure of heterogeneity and should be replaced by tau to quantify heterogeneity of true effect sizes. Additionally, tau instead of tau^2 would be a more informative value to be reported in the tables.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., and Rothstein, H. R. (2017) Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Res. Syn. Meth.*, 8: 5– 18. doi: 10.1002/jrsm.1230.

With regard to the discussion, we personally believe that the results might have warranted an even stronger conclusion: The evidence seems to be strong enough to discourage running any further studies in the area. However, we also noticed that the discussion was very strongly worded and might be perceived as incendiary by proponents of social priming. We are concerned that the tone of the discussion might incite further hostilities between the camp of original authors invested in the theories and the replicators. Furthermore, in our view, the argument against the “lack of skill” hypothesis that might be brought forward by proponents of social priming was not well supported. The authors did not provide evidence that the (majority of) replication studies were run by well-trained researchers. Additionally, even highly skilled scientists may find non-significant results purely based on chance, even if the effect exists. Therefore, a maybe more reasonable recommendation would be to include original authors in preregistered replication efforts, for example, in the form of adversarial collaboration efforts (see, e.g., Vohs et al., 2021; https://psyarxiv.com/e497p).