Response to Comment on “Estimating the reproducibility of psychological science”


Gilbert et al. conclude that evidence from the Open Science Collaboration’s Reproducibility Project: Psychology indicates high reproducibility, given the study methodology. Their very optimistic assessment is limited by statistical misconceptions and by causal inferences from selectively interpreted, correlational data. Using the Reproducibility Project: Psychology data, both optimistic and pessimistic conclusions about reproducibility are possible, and neither are yet warranted.

Across multiple indicators of reproducibility, the Open Science Collaboration’s (OSC2015) first study (1) observed that the original result was replicated in ~40 of 100 studies sampled from three journals. Gilbert et al. (2) conclude that the reproducibility rate is, in fact, as high as could be expected, given the study methodology. We agree with them that both statistical misconceptions and replication studies and statistical power emerge as key factors. We also agree with them that the six methodological differences discussed in the study were present in the original studies. However, these differences did not have a statistically significant impact on the results.

First, Gilbert et al.’s estimates are based on pairwise comparisons between the replication rate in the original studies, while we focus on the overall replication rate across all studies within the project. Second, Gilbert et al. apply the by-site variability from the original studies, while we focus on the overall variability across all studies. Third, Gilbert et al. apply the CI measure, while we focus on the overall replication rate across all studies. Fourth, Gilbert et al. apply the CI measure, while we focus on the overall replication rate across all studies. Fifth, Gilbert et al. apply the CI measure, while we focus on the overall replication rate across all studies. Sixth, Gilbert et al. apply the CI measure, while we focus on the overall replication rate across all studies.

We agree with Gilbert et al. that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. However, we believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. We also believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology.

We agree with Gilbert et al. that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. However, we believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. We also believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology.

We agree with Gilbert et al. that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. However, we believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. We also believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology.

We agree with Gilbert et al. that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. However, we believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. We also believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology.

We agree with Gilbert et al. that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. However, we believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. We also believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology.

We agree with Gilbert et al. that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. However, we believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. We also believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology.

We agree with Gilbert et al. that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. However, we believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. We also believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology.

We agree with Gilbert et al. that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. However, we believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. We also believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology.

We agree with Gilbert et al. that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. However, we believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology. We also believe that the results of the Reproducibility Project: Psychology indicate high reproducibility, given the study methodology.
by Gilbert et al., and a fourth (the racial bias study from America replicated in Italy) was replicated successfully. Gilbert et al. also supposed that non-endorsement of protocols by the original authors was evidence of critical methodological differences. Then they showed that replications that were endorsed by the original authors were more likely to be replicated than those not endorsed (nonendorsed studies included 18 original authors not responding and 11 voicing concerns). In fact, OSC2015 tested whether rated similarity of the replication and original study was correlated with replication success and observed weak relationships across reproducibility indicators ($r = 0.015$ with $P < 0.05$ criterion; supplementary information, p. 67; https://osf.io/k9rnd).

Further, there is an alternative explanation for the correlation between endorsement and replication success; authors who were less confident of their study’s robustness may have been less likely to endorse the replications. Consistent with the alternative account, prediction markets administered on OSC2015 studies showed that it is possible to predict replication failure in advance based on a brief description of the original finding (7). Finally, Gilbert et al. ignored correlational evidence in OSC2015 countering their interpretation, such as evidence that surprising or more underpowered research designs (e.g., interaction tests) were less likely to be replicated. In sum, Gilbert et al. made a causal interpretation for OSC2015’s reproducibility with selective interpretation of correlational data. A constructive step forward would be revising the previously non-endorsed protocols to see if they can achieve endorsement and then conducting replications with the updated protocols to see if reproducibility rates improve.

More generally, there is no such thing as exact replication (8–10). All replications differ in innumerable ways from original studies. They are conducted in different facilities, in different weather, with different experimenters, with different computers and displays, in different languages, at different points in history, and so on. What counts as a replication involves theoretical assessments of the many differences expected to moderate a phenomenon. OSC2015 defined (direct) replication as “the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding.” When results do not differ, it offers some evidence that the finding is generalizable. OSC2015 provides initial, not definitive, evidence—just like the original studies it replicated.

REFERENCES AND NOTES

ACKNOWLEDGMENTS
Preparation of this response was supported by grants from the Laura and John Arnold Foundation and the John Templeton Foundation.

2 December 2015; accepted 28 January 2016
10.1126/science.aad9163
Response to Comment on "Estimating the reproducibility of psychological science"


Science 351 (6277), 1037.
DOI: 10.1126/science.aad9163