TECHNICAL COMMENT

PSYCHOLOGY

Comment on "Estimating the reproducibility of psychological science"

Daniel T. Gilbert,[1]*† Gary King,[1] Stephen Pettigrew,[1] Timothy D. Wilson[2]

A paper from the Open Science Collaboration (Research Articles, 28 August 2015, aac4716) attempting to replicate 100 published studies suggests that the reproducibility of psychological science is surprisingly low. We show that this article contains three statistical errors and provides no support for such a conclusion. Indeed, the data are consistent with the opposite conclusion, namely, that the reproducibility of psychological science is quite high.

The replication of empirical research is a critical component of the scientific process, and attempts to assess and improve the reproducibility of science are important. The Open Science Collaboration (OSC) (*1*) conducted "a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science" by attempting to replicate 100 original studies that had been published in one of three top-tier psychology journals in 2008. Depending on the criterion used, only 36 to 47% of the original studies were successfully replicated, which led many to conclude that there is a "replication crisis" in psychological science (*2*). Here, we show that when these results are corrected for error, power, and bias, they provide no support for this conclusion. In fact, the data are consistent with the opposite conclusion, namely, that the reproducibility of psychological science is quite high.

First, we will discuss the issue of error. If an original study reports a true effect, and if a replication study uses the original procedures with a new sample of subjects drawn from the original population, the replication study will sometimes fail to replicate the original effect because of sampling error alone. If all 100 of the original studies examined by OSC had reported true effects, then sampling error alone should cause 5% of the replication studies to "fail" by producing results that fall outside the 95% confidence interval of the original study and 8% to "fail" by producing results that are not also statistically significant (with the same sign). OSC used the latter figure as the benchmark to which the actual replication failure rate in their data was compared. Neither of these figures provides an appropriate benchmark, however, because both assume that sampling error is the only source of error in the data. In other words, these benchmarks assume that the one and only way in which OSC's replication studies differed from the original studies is that they drew new samples from the original population. In fact, many of OSC's replication studies differed from the original studies in other ways as well.

For example, many of OSC's replication studies drew their samples from different populations than the original studies did. An original study that measured American's attitudes toward African-Americans (*3*) was replicated with Italians, who do not share the same stereotypes; an original study that asked college students to imagine being called on by a professor (*4*) was replicated with participants who had never been to college; and an original study that asked students who commute to school to choose between apartments that were short and long drives from campus (*5*) was replicated with students who do not commute to school. What's more, many of OSC's replication studies used procedures that differed from the original study's procedures in substantial ways: An original study that asked Israelis to imagine the consequences of military service (*6*) was replicated by asking Americans to imagine the consequences of a honeymoon; an original study that gave younger children the difficult task of locating targets on a large screen (*7*) was replicated by giving older children the easier task of locating targets on a small screen; an original study that showed how a change in the wording of a charitable appeal sent by mail to Koreans could boost response rates (*8*) was replicated by sending 771,408 e-mail messages to people all over the world (which produced a response rate of essentially zero in all conditions).

All of these infidelities are potential sources of random error that the OSC's benchmark did not take into account. So how many of their replication studies should we expect to have failed by chance alone? Making this estimate requires having data from multiple replications of the same original study. Although OSC did not collect such data, the corresponding author of OSC, Brian Nosek, referred us to another of his projects that did. The "Many Labs" project (MLP) (*9*) involved 36 independent laboratories that attempted to replicate each of 16 original psychology studies, resulting in 574 replication studies. These replication studies, like OSC's replication studies, did not always use original populations and procedures, so their data allow us to estimate the amount of error that sampling and infidelity together introduce. To make this estimate, we simply treated each of the studies reported by MLP as an "original effect" and then counted how many of the remaining "replications" of that particular study observed that original effect. This analysis revealed that when infidelities were allowed, only 65.5% of the "replication effects" fell within the confidence intervals of the "original effects." Applying this estimate to OSC's data produces a sobering conclusion: If every one of the 100 original studies that OSC attempted to replicate had described a true effect, then more than 34 of their replication studies should have failed by chance alone. [All information and code necessary to replicate our results are archived in Dataverse (*10*).] The bottom line is that OSC allowed considerable infidelities that introduced random error and decreased the replication rate but then compared their results to a benchmark that did not take this error into account.

Second, we will discuss the issue of power. OSC attempted to replicate each of 100 studies just once, and that attempt produced an unsettling result: Only 47% of the original studies were successfully replicated (i.e., produced effects that fell within the confidence interval of the original study). In contrast, MLP attempted to replicate each of its studies 35 or 36 times and then pooled the data. MLP's much more powerful method produced a much more heartening result: A full 85% of the original studies were successfully replicated. What would have happened to MLP's heartening result if they had used OSC's method? Of MLP's 574 replication studies, only 195 produced effects that fell within the confidence interval of the original, published study. In other words, if MLP had used OSC's method, they would have reported an unsettling replication rate of 34% rather than the heartening 85% they actually reported. (A similar result occurs when we limit our analysis to those MLP replication studies that had sample sizes at least as large as the original studies.) Clearly, OSC used a method that severely underestimates the actual rate of replication.

Third, we will discuss the issue of bias. The foregoing analyses generously assume that infidelities are a source of random error that are equally likely to increase or decrease the likelihood of successful replication. Is this assumption true, or were the infidelities in OSC's replication studies more likely to decrease than to increase the likelihood of successful replication? Answering this question requires an indicator of the fidelity of each replication study, which OSC attempted to provide. Before conducting each replication study, OSC asked the authors of the original study whether they endorsed the methodological

[1]Harvard University, Cambridge, MA, USA. [2]University of Virginia, Charlottesville, VA, USA.
*Corresponding author. E-mail: gilbert@wjh.harvard.edu
†Authors are listed alphabetically.

protocol for the to-be-attempted replication. Only 69% of the original authors did. Although endorsement is an imperfect indicator that may overestimate the fidelity of a replication study (e.g., some of the original authors may have knowingly endorsed low-fidelity protocols and others may have discovered that the replication studies were low fidelity only after they were completed) or may underestimate the fidelity of a replication study (e.g., endorsement decisions may be influenced by original authors' suspicions about the weakness of their studies rather than by the fidelity of the replication protocol), it is nonetheless the best indicator of fidelity in OSC's data. So what does that indicator indicate?

When we compared the replication rates of the endorsed and unendorsed protocols, we discovered that the endorsed protocols were nearly four times as likely to produce a successful replication (59.7%) as were the unendorsed protocols (15.4%). This strongly suggests that the infidelities did not just introduce random error but instead biased the replication studies toward failure. If OSC had limited their analyses to endorsed studies, they would have found that 59.7% [95% confidence interval (CI): 47.5%, 70.9%] were replicated successfully. In fact, we estimate that if all the replication studies had been high enough in fidelity to earn the endorsement of the original authors, then the rate of successful replication would have been 58.6% (95% CI: 47.0%, 69.5%) when controlling for relevant covariates. Remarkably, the CIs of these estimates actually overlap the 65.5% replication rate that one would expect if every one of the original studies had reported a true effect. Although that seems rather unlikely, OSC's data clearly provide no evidence for a "replication crisis" in psychological science.

We applaud efforts to improve psychological science, many of which have been careful, responsible, and effective (*11*), and we appreciate the effort that went into producing OSC. But metascience is not exempt from the rules of science. OSC used a benchmark that did not take into account the multiple sources of error in their data, used a relatively low-powered design that demonstrably underestimates the true rate of replication, and permitted considerable infidelities that almost certainly biased their replication studies toward failure. As a result, OSC seriously underestimated the reproducibility of psychological science.

REFERENCES

1. Open Science Collaboration, *Science* **349**, aac4716 (2015).
2. B. Carey, Psychology's fears confirmed: Rechecked studies don't hold up. *New York Times* (27 August 2015), p. A1.
3. B. K. Payne, M. A. Burkley, M. B. Stokes, *J. Pers. Soc. Psychol.* **94**, 16–31 (2008).
4. J. L. Risen, T. Gilovich, *J. Pers. Soc. Psychol.* **95**, 293–307 (2008).
5. E. J. Masicampo, R. F. Baumeister, *Psychol. Sci.* **19**, 255–260 (2008).
6. N. Shnabel, A. Nadler, *J. Pers. Soc. Psychol.* **94**, 116–132 (2008).
7. V. LoBue, J. S. DeLoache, *Psychol. Sci.* **19**, 284–289 (2008).
8. M. Koo, A. Fishbach, *J. Pers. Soc. Psychol.* **94**, 183–195 (2008).
9. R. A. Klein *et al.*, *Soc. Psychol.* **45**, 142–152 (2014).
10. D. T. Gilbert, G. King, S. Pettigrew, T. D. Wilson, Replication data for Comment on "Estimating the reproducibility of psychological science," Harvard Dataverse, V1; http://dx.doi.org/10.7910/DVN/5LKVH2 (2016).
11. J. P. Simmons, L. D. Nelson, U. Simonsohn, *Psychol. Sci.* **22**, 1359–1366 (2011).

ACKNOWLEDGMENTS

We acknowledge the support of NSF Grant BCS-1423747 to T.D.W. and D.T.G.

26 October 2015; accepted 28 January 2016
10.1126/science.aad7243

Technical Comment: Supplementary Appendix

Title: Supplementary appendix for Comment on "Estimating the reproducibility of psychological science."

Authors: Daniel T. Gilbert[1], Gary King[1], Stephen Pettigrew[1], Timothy D. Wilson[2] *(1)*

Affiliations: Harvard University[1], University of Virginia[2]

Summary: One way the Open Science Collaboration's recent report (*1,* hereinafter referred to as OSC2015) defines a "successful replication" is as one in which the 95% confidence interval (CI) from the replication captures the point estimate from the original published result (see OSC2015 Table 1, column 10). This does not make statistical sense because the effect that they were trying to replicate was that of the original study, not of the replication. Thus, we inverted the calculation to determine whether the point estimates from the replications were captured within the 95% CI of the original article. Making this switch has a negligible impact on the numbers reported below and changes no conclusions.

If the only change made to the original studies was that new samples were drawn from each of the original populations, we would expect that 95% of effects from the replications would fall inside the CI of the original studies due to chance alone. OSC2015 finds that only 47% studies replicated based on this criterion (OSC2015, Table 1, column 10). However, we know that in addition to drawing new samples, the replication experiments conducted in OSC2015 deviated from the design protocols of the original studies much more than simply drawing a new sample.

We conclude: (a) The percent of studies in OSC2015 that should be expected to fail to replicate by chance alone is not 5%, but at least 34.54%, and probably much higher; (b) If the replication studies in OSC2015 had been more highly powered, the observed replication rate would have been quite high; and (c) If OSC2015 had analyzed only the high fidelity replications that were endorsed by the original authors, the percent of successful replications would have been statistically indistinguishable from 100%.

Details:

The Many Labs Design. In Many Labs project (*2,* hereinafter referred to as ML2014), the authors chose 13 previously published studies from psychological science and had 36 different groups of researchers ("many labs") attempt to replicate each study. Each lab drew a sample from a different population and there were differences between labs in how the samples were

drawn. One of the 13 studies was replicated in four different ways, and a few were replicated in fewer than 36 labs. This resulted in a total of 574 total replication datasets.

Replication information: All information and code necessary to replicate our results are archived in dataverse (3).

1. ERROR

The total uncertainty due to replication could be estimated by having *fully independent* teams replicate the same study, where "fully independent" means absolutely no communication between the teams. Overall, total replication uncertainty is due to (a) identifying the population of interest, (b) sampling subjects from that population, and (c) choosing and implementing the experimental protocols (such as in-person versus online interactions, details about the survey or other measurement instrument, how the treatment was administered, what covariates or experimental conditions were held physically or statistically constant, what statistical estimator was used, etc.). The more uncertainty in replication, the larger the number of studies we should expect to fail to replicate by chance alone.

The replication procedures used in OSC2015 ensured considerable variability from all three sources, and yet, when the authors of OSC2015 computed the number of studies one should expect to fail, they assumed that the only source of variability was item (b) above. They explained that "on the basis of only the average replication power of the 97 original, significant effects [M = 0.92, median (Mdn) = 0.95], we would expect approximately 89 positive results in the replications if all original effects were true and accurately estimated." (OSC2015, page aac4716-4) This estimate of a 92% success rate is the average of the statistical power for the 100 replication studies and represents the expected number of replication point estimates that are statistically significant and in the same direction of the original result. OSC2015 does not provide a similar baseline for the CI replication test from Table 1, column 10, although based on statistical theory we know that 95% of replication estimates should fall within the 95% CI of the original results. Given that the replication protocols were different from those of the original studies, there is strong reason to suspect that these figures severely overestimate the actual number of studies one should expect to fail to replicate by chance alone.

ML2014 replicated the same study 35 or 36 times but had many dependencies across the replications: "We bundled the selected studies together into a brief, easy-to-administer experiment that was delivered to each participating sample [i.e., from each lab] through a single infrastructure (http://projectimplicit.net/). There are many factors that can influence the replicability of an effect such as sample, setting, statistical power, and procedural variations. The present design standardizes procedural characteristics and ensures appropriate statistical power in

order to examine the effects of sample and setting on replicability" (ML2014, page 143). This means that we can use the ML2014 data to estimate the number of studies that would be expected not to replicate on the basis of uncertainty due to uncertainty from (a), (b), but only parts of (c). Thus, a calculation using these data will be more realistic than the one in OSC2015 which is based on (b) alone but will still underestimate the number of failures to replicate that are due to chance.

We first used the CI from one lab's test of one of the 13 studies (and no data from the original studies they attempted to replicate) as a "baseline" (analogous to the situation where this particular replication had been the published result). We treated the tests of that study from the other 35 labs as "replications" of this result. Repeating this for all 574 study-lab combinations, we find that 34.5% of replications generated a point estimate outside the 95% CI of the "published" result. This far exceeds the 5% failure rate we would expect due to (b) alone.

Conclusion 1. OSC2015 concluded that if 100% of the original studies they attempted to replicate had produced true effects, they would expect 92% of them (or 95% based on the metric we are using) to replicate. In fact, using the ML2014 data solely to estimate reliability of the replicators, we should expect only 65.5% of those original studies to be successfully replicated given OSC2015's design. Because this latter estimate excludes many sources of uncertainty due to (c), and because the deviations between the protocols used in the articles replicated by OSC2015 and the single replications we observe are so large, *the percent of original articles containing true effects that one should expect to replicate given the OSC2015 design would probably be lower than 66%.*

2. POWER

The authors of ML2014 assessed replication success for each original published study by pooling together data from all (approximately) 36 labs and conducted one high-powered analysis. For 11 of the 13 articles, this pooled test statistic was statistically significant and in the same direction as the original study, and by most other criteria, 10 or 11 of the 13 studies were successfully replicated. The authors of ML2014 took this as evidence that original results were replicated in 11 of 13 or 85% of the articles. Clearly, they found no evidence of a "replication crisis" in psychological science.

To emulate the setup of OSC2015 using the ML2014 data, we considered each of the 574 replications individually, rather than pooling them. We calculated the 95% CI of the estimate reported in the 13 original, published articles and determined whether the point estimate of the 574 replications fell within this interval. We found that 34.0% of the 574 replications were within the 95% CI reported by the original author, which would seem to suggest even more of a

"replication crisis" than does OSC2015, in which 47% replicated based on this criterion (OSC2015, Table 1, column 10).

This approach mirrors the one taken by OSC2015, where replication sample size was roughly the same as the original published article. When we rerun the above analysis using only the ML2014 replication that had a sample at least as big as the original published article, our conclusions remain the same. Among the highest powered Many Labs replications, just 40.7% reported a result inside the 95% CI of the original study.

Conclusion 2. ML2014, which combined all the data from 36 studies into one pooled study, was very powerful, and on that basis concluded that 11 of 13 or 85% of the original studies were replicated. In other words, there is no evidence of a "replication crisis" in psychological science. By selecting one replication out of the 35 or 36 replications from ML2014, we approximated the procedure used in OSC2015. When we did this, we found that the probability of replication in a single study was only about 34%, which is even worse than the OSC2015 figure of 47%. The implication is that if each of the 100 original studies examined in OSC2015 had been replicated with a more powerful design (or if many studies were conducted and pooled together, as was done in ML2014), an extremely large percentage of the 100 original studies would have replicated. The clear conclusion is that *OSC2015 provides no evidence for a replication crisis in psychological science.*

3. BIAS

Prior to running each replication in the OSC2015 study, each replication team contacted the authors of the original study that they were seeking to replicate (given protocols provided by the managing team). The original authors were asked the extent to which they endorsed the replication protocol prior to running the experiment. Most (69%) endorsed the replication protocol, 8% had concerns about the replication plan based on informed judgment or speculation, 3% had concerns based on published empirical evidence about the constraints of the effect, and 18% did not respond. We asked: What would the replication rate have been if all of the replications in OSC2015 had been high enough in fidelity to have been endorsed by the original authors? Or alternatively, if OSC2015 had only analyzed the 69% of studies which were endorsed by the original authors?

In OSC2015, endorsement was a strong predictor of replication success: Of the replication protocols that were not endorsed by original authors, only 15.4% were successfully replicated, but of the replication protocols that were endorsed by the original authors, a striking 3.9 times more -- or 59.7% -- were successfully replicated. If OSC2015 had only analyzed the replications

that were endorsed by the original authors, their success rate would have increased from 47% to 59.7% (95% CI: [47.7%, 70.4%]).

We also estimated a logistic regression model of replication success on the degree of endorsement by the original authors, operationalized as a categorical variable with 4 levels. We used these regression results to generate predicted probabilities of successful replication in the counterfactual world in which all studies were endorsed. In this world, in which all replications were of high enough fidelity to have been endorsed by the original authors, our model predicted that the replication rate would have been 58.6% (95% CI: [47.0%, 69.5%]), controlling for covariates (the citation count of the original paper, the discipline/subfield of the research, and pre-replication assessments of whether high level methodological expertise would be required and whether the original paper's design had a high probability of expectancy bias). These confidence intervals overlap the (lower than) 65.5% rate of successful replication that we would expect by chance if all of the original studies had produced true effects and all of the sources of uncertainty had been considered.

This last analysis is an extrapolation to a counterfactual situation and by definition more model dependent. The uncertainties generated by this model dependence do not mean this model should not be run; at present, it is the only way to estimate the quantity of interest that OSC2015 sought to estimate. Moreover, the uncertainty exists whether or not the model is run, including if, like OSC2015, one were to ignore the mistakes due to the replication infidelities.

Conclusion 3. *If OSC2015 had conducted only replications that were of high enough fidelity to be endorsed by original authors, the percent of successful replications they observed would have been statistically indistinguishable from the percent one would expect if all of the original effects had been true.*

References

1. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science.* **349**, aac4716–1–aac4716–8 (2015).
2. R. A. Klein *et al*, Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology* **45**(3): 142-152 (2014).
3. Gilbert, Daniel; King, Gary; Pettigrew, Stephen; Wilson, Timothy, 2016, "Replication Data for: Comment on `Estimating the reproducibility of psychological science.'", http://dx.doi.org/10.7910/DVN/5LKVH2, Harvard Dataverse, V1 [UNF:6:2+oJQsG7lz6j6Kqaj/bF+g==]