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