TECHNICAL RESPONSE

PSYCHOLOGY

Response to Comment on "Estimating the reproducibility of psychological science"

Christopher J. Anderson,[1]* Štěpán Bahník,[2] Michael Barnett-Cowan,[3] Frank A. Bosco,[4] Jesse Chandler,[5,6] Christopher R. Chartier,[7] Felix Cheung,[8] Cody D. Christopherson,[9] Andreas Cordes,[10] Edward J. Cremata,[11] Nicolas Della Penna,[12] Vivien Estel,[13] Anna Fedor,[14] Stanka A. Fitneva,[15] Michael C. Frank,[16] James A. Grange,[17] Joshua K. Hartshorne,[18] Fred Hasselman,[19] Felix Henninger,[20] Marije van der Hulst,[21] Kai J. Jonas,[22] Calvin K. Lai,[23] Carmel A. Levitan,[24] Jeremy K. Miller,[25] Katherine S. Moore,[26] Johannes M. Meixner,[27] Marcus R. Munafò,[28] Koen I. Neijenhuijs,[29] Gustav Nilsonne,[30] Brian A. Nosek,[31,32]† Franziska Plessow,[33] Jason M. Prenoveau,[34] Ashley A. Ricker,[35] Kathleen Schmidt,[36] Jeffrey R. Spies,[31,32] Stefan Stieger,[37] Nina Strohminger,[38] Gavin B. Sullivan,[39] Robbie C. M. van Aert,[40] Marcel A. L. M. van Assen,[40,41] Wolf Vanpaemel,[42] Michelangelo Vianello,[43] Martin Voracek,[44] Kellylynn Zuni[45]

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 (*1*) (OSC2015) 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 methodological differences between original and replication studies and statistical power affect reproducibility, but their very optimistic assessment is based on statistical misconceptions and selective interpretation of correlational data.

Gilbert *et al.* focused on a variation of one of OSC2015's five measures of reproducibility: how often the confidence interval (CI) of the original study contains the effect size estimate of the replication study. They misstated that the expected replication rate assuming only sampling error is 95%, which is true only if both studies estimate the same population effect size and the replication has infinite sample size (*3*, *4*). OSC2015 replications did not have infinite sample size. In fact, the expected replication rate was 78.5% using OSC2015's CI measure (see OSC2015's supplementary information, pp. 56 and 76; https://osf.io/k9rnd). By this measure, the actual replication rate was only 47.4%, suggesting the influence of factors other than sampling error alone.

Within another large replication study, "Many Labs" (*5*) (ML2014), Gilbert *et al.* found that 65.5% of ML2014 studies would be within the CIs of other ML2014 studies of the same phenomenon and concluded that this reflects the maximum reproducibility rate for OSC2015. Their analysis using ML2014 is misleading and does not apply to estimating reproducibility with OSC2015's data for a number of reasons.

First, Gilbert *et al.*'s estimates are based on pairwise comparisons between all of the replications within ML2014. As such, for roughly half of their failures to replicate, "replications" had larger effect sizes than "original studies," whereas just 5% of OSC2015 replications had replication CIs exceeding the original study effect sizes.

Second, Gilbert *et al.* apply the by-site variability in ML2014 to OSC2015's findings, thereby arriving at higher estimates of reproducibility. However, ML2014's primary finding was that by-site variability was highest for the largest (replicable) effects and lowest for the smallest (nonreplicable) effects. If ML2014's primary finding is generalizable, then Gilbert *et al.*'s analysis may leverage by-site variability in ML2014's larger effects to exaggerate the effect of by-site variability on OSC2015's nonreproduced smaller effects, thus overestimating reproducibility.

Third, Gilbert *et al.* use ML2014's 85% replication rate (after aggregating across all 6344 participants) to argue that reproducibility is high when extremely high power is used. This interpretation is based on ML2014's small, ad hoc sample of classic and new findings, as opposed to OSC2015's effort to examine a more representative sample of studies in high-impact journals. Had Gilbert *et al.* selected the similar Many Labs 3 study (*6*) instead of ML2014, they would have arrived at a more pessimistic conclusion: a 30% overall replication success rate with a multisite, very high-powered design.

That said, Gilbert *et al.*'s analysis demonstrates that differences between laboratories and sample populations reduce reproducibility according to the CI measure. Also, some true effects may exist even among nonsignificant replications (our additional analysis finding evidence for these effects is available at https://osf.io/smjge). True effects can fail to be detected because power calculations for replication studies are based on effect sizes in original studies. As OSC2015 demonstrates, original study effect sizes are likely inflated due to publication bias. Unfortunately, Gilbert *et al.*'s focus on the CI measure of reproducibility neither addresses nor can account for the facts that the OSC2015 replication effect sizes were about half the size of the original studies on average, and 83% of replications elicited smaller effect sizes than the original studies. The combined results of OSC2015's five indicators of reproducibility suggest that, even if true, most effects are likely to be smaller than the original results suggest.

Gilbert *et al.* attribute some of the failures to replicate to "low-fidelity protocols" with methodological differences relative to the original, for which they provide six examples. In fact, the original authors recommended or endorsed three of the six methodological differences discussed

[1]Russell Sage College, Troy, NY, USA. [2]University of Würzburg, Würzburg, Germany. [3]University of Waterloo, Waterloo, Ontario, Canada. [4]Virginia Commonwealth University, Richmond, VA, USA. [5]University of Michigan, Ann Arbor, MI 48104, USA. [6]Mathematica Policy Research, Washington, DC, USA. [7]Ashland University, Ashland, OH, USA. [8]Michigan State University, East Lansing, MI, USA. [9]Southern Oregon University, Ashland, OR, USA. [10]University of Göttingen, Institute for Psychology, Göttingen, Germany. [11]University of Southern California, Los Angeles, CA, USA. [12]Australian National University, Canberra, Australia. [13]Technische Universität Braunschweig, Braunschweig, Germany. [14]Parmenides Stiftung, Munich, Germany. [15]Queen's University, Kingston, Ontario, Canada. [16]Stanford University, Stanford, CA, USA. [17]Keele University, Keele, Staffordshire, UK. [18]Boston College, Chestnut Hill, MA, USA. [19]Radboud University Nijmegen, Nijmegen, Netherlands. [20]University of Koblenz-Landau, Landau, Germany. [21]Erasmus Medical Center, Rotterdam, Netherlands. [22]University of Amsterdam, Amsterdam, Netherlands. [23]Harvard University, Cambridge, MA, USA. [24]Occidental College, Los Angeles, CA, USA. [25]Willamette University, Salem, OR, USA. [26]Arcadia University, Glenside, PA, USA. [27]University of Potsdam, Potsdam, Germany. [28]University of Bristol, Bristol, UK. [29]Vrije Universiteit Amsterdam, Amsterdam, Netherlands. [30]Karolinska Institutet, Stockholm University, Stockholm, Sweden. [31]Center for Open Science, Charlottesville, VA, USA. [32]University of Virginia, Charlottesville, VA, USA. [33]Harvard Medical School, Boston, MA, USA. [34]Loyola University, Baltimore, MD, USA. [35]University of California, Riverside, CA, USA. [36]Wesleyan University, Middletown, CT, USA. [37]University of Konstanz, Konstanz, Germany. [38]Yale University, New Haven, CT, USA. [39]Coventry University, Coventry, UK. [40]Tilburg University, Tilburg, Netherlands. [41]Utrecht University, Utrecht, Netherlands. [42]University of Leuven, Leuven, Belgium. [43]University of Padova, Padova, Italy. [44]University of Vienna, Vienna, Austria. [45]Adams State University, Alamosa, CO, USA. *Authors are listed alphabetically. †Corresponding author. E-mail: nosek@virginia.edu

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 nonendorsement 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 (e.g., $r = 0.015$ 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 nonendorsed 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 differ, it offers an opportunity for hypothesis generation and then testing to determine why. 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

1. Open Science Collaboration, *Science* **349**, aac4716 (2015).
2. D. T. Gilbert, G. King, S. Pettigrew, T. D. Wilson, *Science* **351**, xxx (2016).
3. G. Cumming, R. Maillardet, *Psychol. Methods* **11**, 217–227 (2006).
4. G. Cumming, J. Williams, F. Fidler, *Underst. Stat.* **3**, 299–311 (2004).
5. R. A. Klein *et al.*, *Soc. Psychol.* **45**, 142–152 (2014).
6. C. R. Ebersole *et al.*, *J. Exp. Soc. Psychol.* **65** (2016); https://osf.io/csygd.
7. A. Dreber *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15343–15347 (2015).
8. B. A. Nosek, D. Lakens, *Soc. Psychol.* **45**, 137–141 (2014).
9. Open Science Collaboration, *Perspect. Psychol. Sci.* **7**, 657–660 (2012).
10. S. Schmidt, *Rev. Gen. Psychol.* **13**, 90–100 (2009).

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