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

Anderson, C. J. , Bahník, S. , Barnett-Cowan, M. , Bosco, F. A. , Chandler, J. , Chartier, C. R. , Cheung, F. , Christopherson, C. D. , Cordes, A. , Cremata, E. J. , Della Penna, N. , Estel, V. , Fedor, A. , Fitneva, S. A. , Frank, M. C. , Grange, J. A. , Hartshorne, J. K. , Hasselman, F. , Henninger, F. , van der Hulst, M. , Jonas, K. J. , Lai, C. K. , Levitan, C. A. , Miller, J. K. , Moore, K. S. , Meixner, J. M. , Munafò, M. R. , Neijenhuijs, K. I. , Nilsonne, G. , Nosek, B. A. , Plessow, F. , Prenoveau, J. M. , Ricker, A. A. , Schmidt, K. , Spies, J. R. , Stieger, S. , Strohminger, N. , Sullivan, G.B. , van Aert, R. C. M. , van Assen, M. A. L. M. , Vanpaemel, W. , Vianello, M. , Voracek, M. and Zuni, K.

Original citation & hyperlink:
Anderson, C. J. , Bahník, S. , Barnett-Cowan, M. , Bosco, F. A. , Chandler, J. , Chartier, C. R. , Cheung, F. , Christopherson, C. D. , Cordes, A. , Cremata, E. J. , Della Penna, N. , Estel, V. , Fedor, A. , Fitneva, S. A. , Frank, M. C. , Grange, J. A. , Hartshorne, J. K. , Hasselman, F. , Henninger, F. , van der Hulst, M. , Jonas, K. J. , Lai, C. K. , Levitan, C. A. , Miller, J. K. , Moore, K. S. , Meixner, J. M. , Munafò, M. R. , Neijenhuijs, K. I. , Nilsonne, G. , Nosek, B. A. , Plessow, F. , Prenoveau, J. M. , Ricker, A. A. , Schmidt, K. , Spies, J. R. , Stieger, S. , Strohminger, N. , Sullivan, G.B. , van Aert, R. C. M. , van Assen, M. A. L. M. , Vanpaemel, W. , Vianello, M. , Voracek, M. and Zuni, K. (2016) Response to Comment on “Estimating the reproducibility of psychological science”. Science, volume 251 (6277): 1037

URL http://dx.doi.org/10.1126/science.aad9163
DOI 10.1126/science.aad9163
ISSN 0036-8075
ESSN 1095-9203
Publisher American Association for the Advancement of Science

Response to a comment on “Estimating the Reproducibility of Psychological Science”

Abstract

Gilbert, King, Pettigrew, and Wilson conclude that evidence from the 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.

Main Body

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. (GKPW) conclude that the reproducibility rate is, in fact, as high as could be expected given the study methodology. We agree with GKPW 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.

GKPW 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. GKPW 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 (*2,3*). 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 SI p. 56, 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” (*4*, ML2014), GKPW found that 65.5% of ML2014 studies would be within the confidence intervals 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, GKPW’s estimates are based on pairwise comparisons between all of the replications within ML2014. As such, for roughly half of GKPW’s *failures to replicate*, “replications” had *larger* effect sizes than “original studies” whereas just 5% of OSC2015 replications had replication CI’s exceeding the original study effect sizes.

Second, GKPW 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 (non-replicable) effects. If ML2014’s primary finding is generalizable, then GKPW’s analysis may leverage by-site variability in ML2014’s larger effects to exaggerate the impact of by-site variability on OSC2015’s non-reproduced smaller effects, thus overestimating reproducibility.

Third, GKPW 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 GKPW selected the similar Many Labs 3 study (*5*) they would have arrived at a more pessimistic conclusion: a 30% overall replication success rate with a multi-site, very high-powered design.

That said, GKPW’s analysis demonstrates that differences between labs and sample populations reduce reproducibility according to the CI measure. Also, some true effects may exist even among non-significant 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, GKPW'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.

GKPW attribute some of the failures to replicate to "low-fidelity replications" 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 by GKPW, and a fourth (the racial bias study from America replicated in Italy) replicated successfully. GKPW also supposed that non-endorsement of protocols by original authors was evidence of critical methodological differences. Then they showed that replications that were endorsed by the original authors were more likely to replicate than those not endorsed (non-endorsed 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 = .015 with "p < .05" criterion, SI, 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 *(6)*. Finally, GKPW 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 replicate. In sum, GKPW 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" *(7-9)*. 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. Estimating the reproducibility of psychological science. *Science* **349**, 943 (2015).

2. G Cumming, R Maillardet. Confidence intervals and replication: where will the next mean fall?. *Psychol. methods,* **11**, 217-227 (2006).

3. G Cumming, J Williams, F Fidler. Replication and researchers' understanding of confidence intervals and standard error bars. *Underst. Stat.,* **3**, 299-311 (2004).

4. RA Klein, KA Ratliff, M Vianello, RB Adams Jr, Š Bahník, *et al.* Investigating variation in replicability. *Soc. Psychol*. **45**, 142-152 (2014).

5. CR Ebersole, OE Atherton, AL Belanger, HM Skulborstad, RB Adams, *et al.* Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* (2016).

6. A Dreber, T Pfeiffer, J Almenberg, S Isaksson, B Wilson, *et al.* Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl. Acad. Sci.U.S.A.* (2015), doi:10.1073/pnas.1516179112.

7. BA Nosek, D Lakens. Registered reports: A method to increase the credibility of published results. *Soc. Psychol.* **45**, 137-141 (2014).

8. Open Science Collaboration. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. Psychol. Sci.* **7**, 657-660 (2012).

9. S Schmidt, Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* **13**, 90-100 (2009).

------ Acknowledgements -----

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

------------- Authors (alphabetical order) -------------

Name	**Email**	**Institution**
Christopher J. Anderson	anderc@sage.edu	Russell Sage College
Štěpán Bahník	bahniks@seznam.cz	University of Würzburg
Michael Barnett-Cowan	mbc@uwaterloo.ca	University of Waterloo
Frank A. Bosco	fabosco@vcu.edu	Virginia Commonwealth University
Jesse Chandler	jjchandl@umich.edu	University of Michigan; Mathematica Policy Research
Christopher R. Chartier	cchartie@ashland.edu	Ashland University
Felix Cheung	felixckc@msu.edu	Michigan State University
Cody D. Christopherson	christoc1@sou.edu	Southern Oregon University
Andreas Cordes	acordes@uni-goettingen.de	University of Göttingen, Institute for Psychology
Edward J. Cremata	cremata@usc.edu	University of Southern California
Nicolas Della Penna	nicolas.della-penna@anu.edu.au	Australian National University
Vivien Estel	v.estel@tu-bs.de	Technische Universität Braunschweig
Anna Fedor	fedoranna@gmail.com	Parmenides Stiftung
Stanka A. Fitneva	fitneva@queensu.ca	Queen's University
Michael C. Frank	mcfrank@stanford.edu	Stanford University
James A. Grange	j.a.grange@keele.ac.uk	Keele University
Joshua K Hartshorne	joshua.hartshorne@bc.edu	Boston College
Fred Hasselman	f.hasselman@pwo.ru.nl	Radboud University Nijmegen
Felix Henninger	mailbox@felixhenninger.com	University of Koblenz-Landau
Marije van der Hulst	m.vanderhulst@erasmusmc.nl	Erasmus Medical Center
Kai J. Jonas	k.j.jonas@uva.nl	University of Amsterdam
Calvin K. Lai	cklai4@gmail.com	Harvard University
Carmel A. Levitan	levitan@oxy.edu	Occidental College
Jeremy K. Miller	millerj@willamette.edu	Willamette University
Katherine S. Moore	moorek@arcadia.edu	Arcadia University
Johannes M. Meixner	johannes.meixner@uni-potsdam.de	University of Potsdam, Germany
Marcus R. Munafò	marcus.munafo@bristol.ac.uk	University of Bristol
Koen I. Neijenhuijs	k.i.neijenhuijs@vu.nl	VU University Amsterdam
Gustav Nilsonne	gustav.nilsonne@ki.se	Karolinska Institutet, Stockholm University
Brian A. Nosek	nosek@virginia.edu	Center for Open Science; University of Virginia
Franziska Plessow	fplessow@mgh.harvard.edu	Harvard Medical School
Jason M. Prenoveau	jmprenoveau@loyola.edu	Loyola University Maryland
Ashley A. Ricker	ashley.ricker@email.ucr.edu	University of California, Riverside
Kathleen Schmidt	kschmidt@wesleyan.edu	Wesleyan University

Jeffrey R. Spies	jeff@cos.io	Center for Open Science; University of Virginia
Stefan Stieger	stefan.stieger@uni-konstanz.de	University of Konstanz
Nina Strohminger	nina.strohminger@yale.edu	Yale University
Gavin B. Sullivan	gavin.sullivan@coventry.ac.uk	Coventry University
Robbie C.M. van Aert	R.C.M.vanAert@tilburguniversity.edu	Tilburg University
Marcel A.L.M. van Assen	m.a.l.m.vanassen@uvt.nl	Tilburg University, Utrecht University
Wolf Vanpaemel	wolf.vanpaemel@ppw.kuleuven.be	University of Leuven
Michelangelo Vianello	michelangelo.vianello@unipd.it	University of Padova
Martin Voracek	martin.voracek@univie.ac.at	University of Vienna
Kellylynn Zuni	zuni.kellylynn@gmail.com	Adams State University