Best practices for interpreting large-scale replications

To the Editor — I appreciate the opportunity to comment on the Social Science Replication Project (SSRP)¹. I thank the team for their attention to the important issue of reproducibility in social science. The work being done on this issue, including the current project, is vital for improving the health of scientific fields. The project team replicated the first study from our paper², producing mixed evidence for the central effect: the stage 1 attempt replicated, but the stage 2 attempt did not; the meta-analytic effect showed evidence of a significant effect, though again, certain complementary analytic procedures did not. I believe the team endeavoured to reproduce the original study procedure as exactly as possible, despite some hiccups along the way. Most importantly, a second replication study was run in error (featuring changes in sample characteristics and a larger number of participants excluded). This error spoke to a more general concern I had about the SSRP report.

Specifically, there are many ways to evaluate reproducibility of findings. The current method isolated individual studies and reproduced them once if replication was achieved, twice otherwise. My concern is that conclusions drawn from this approach may sometimes extend beyond its inherent limitations. To be sure, the report provides an illuminating look at research published in two top scientific journals, and the prediction market and complementary replicability indicators are extremely useful for understanding (and learning about) how

we can evaluate effects. But does focusing on one or two replications of a variety of studies provide that much value?

Consider the unique case of how our study was reproduced. During the first attempt to replicate our study, the analyses contained an error, which led the replication team to incorrectly conclude that the replication had failed. They therefore conducted a stage 2 study, at which point the original error became clear. If the stage 1 replication analyses had originally been conducted correctly, the team would have concluded that the effect replicates and no second replication study would have been run. Yet, this mistaken second study found no sufficient replication evidence. Presumably, a similar replicated/failed-to-replicate pattern could characterize all other investigations that terminated at stage 1 (11 of the studies). We cannot know using a procedure like the one here. If we want to gain meaningful information about the reproducibility of specific effects, my guess is that the current procedure will not help much. A better approach may be for many groups to reproduce a single study many times³. In the meantime, given variation in reproducibility across stages for other studies in the SSRP, it may be reasonable to give more weight to the meta-analytic effects than single replication failures (though the team's point about publication bias is certainly relevant).

I raise this concern primarily because I increasingly hear problematic conclusions

drawn by colleagues, students and the public following projects like the SSRP. People sometimes presume that one failure to replicate is the final word on an idea instead of a point made within an ongoing conversation. My suspicion is that highlighting a conclusion such as 'X% of studies failed to replicate' following only one or two attempts unfortunately contributes to this problem.

Of course, the SSRP team is surely aware of such issues. Disagreements about approach are bound to emerge as the field sorts through the best practices needed to improve our science, and the current project is an important step along this path. □

Joshua M. Ackerman

University of Michigan, Ann Arbor, MI, USA.
e-mail: joshack@umich.edu

Published online: 20 September 2018
<https://doi.org/10.1038/s41562-018-0447-8>

References

1. Camerer, C. F. et al. *Nat. Hum. Behav.* <https://doi.org/10.1038/s41562-018-0399-z> (2018).
2. Ackerman, J. M., Nocera, C. C. & Bargh, J. A. *Science* **328**, 1712–1715 (2010).
3. Hagger, M. S. et al. *Perspect. Psychol. Sci.* **11**, 546–573 (2016).

Acknowledgements

The comments provided are entirely the sole author's.

Competing interests

The author declares no competing interests.