# An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science

Open Science Collaboration[1]

Authors' note: Correspondence can be addressed to nosek@virginia.edu. The authors had no financial conflict of interest in the preparation of this article.

Word count = 1482 without references

[1] Anita Alexander, University of Virginia; Michael Barnett-Cowan, The Brain and Mind Institute, Western University, Canada; Elizabeth Bartmess, University of California, San Francisco; Frank A. Bosco, Marshall University; Mark Brandt, Tilburg University; Joshua Carp , University of Michigan; Jesse J. Chandler, Princeton University; Russ Clay, University of Richmond; Hayley Cleary, Virginia Commonwealth University; Michael Cohn, University of California, San Francisco; Giulio Costantini, University of Milan - Bicocca; Jamie DeCoster, University of Virginia; Elizabeth Dunn, University of British Columbia; Casey Eggleston, University of Virginia; Vivien Estel, University of Erfurt; Frank J. Farach, University of Washington; Jenelle Feather, M.I.T.; Susann Fiedler, Max Planck Institute for Research on Collective Goods; James G. Field, Marshall University; Joshua D. Foster, University of South Alabama; Michael Frank, Stanford University; Rebecca S. Frazier, University of Virginia; Heather M. Fuchs, University of Cologne; Jeff Galak, Carnegie Mellon University; Elisa Maria Galliani, University of Padova; Sara García, Universidad Nacional de Asunción; Elise M. Giammanco, University of Virginia; Elizabeth A. Gilbert, University of Virginia; Roger Giner-Sorolla, University of Kent; Lars Goellner, University of Erfurt; Jin X. Goh, Northeastern University; R. Justin Goss, University of Texas at San Antonio; Jesse Graham, University of Southern California; James A. Grange, Keele University; Jeremy R. Gray, Michigan State University; Sarah Gripshover, Stanford University; Joshua Hartshorne, M.I.T.; Timothy B. Hayes, University of Southern California; Georg Jahn, University of Greifswald; Kate Johnson, University of Southern California; William Johnston, M.I.T.; Jennifer A. Joy-Gaba, Virginia Commonwealth University; Calvin K. Lai, University of Virginia; Daniel Lakens, Eindhoven University of Technology; Kristin Lane, Bard College; Etienne P. LeBel, University of Western Ontario; Minha Lee, University of Virginia; Kristi Lemm, Western Washington University; Sean Mackinnon, Dalhousie University; Michael May, University of Bonn; Katherine Moore, Elmhurst College; Matt Motyl, University of Virginia; Stephanie M. Müller, University of Erfurt; Marcus Munafo, University of Bristol; Brian A. Nosek, University of Virginia; Catherine Olsson, M.I.T.; Dave Paunesku, Stanford University; Marco Perugini, University of Milan - Bicocca; Michael Pitts, Reed College; Kate Ratliff, University of Florida; Frank Renkewitz, University of Erfurt; Abraham M. Rutchick, California State University, Northridge; Gillian Sandstrom, University of British Columbia; Rebecca Saxe, M.I.T.; Dylan Selterman, University of Maryland; William Simpson, University of Virginia; Colin Tucker Smith, University of Florida; Jeffrey R. Spies, University of Virginia; Nina Strohminger, Duke University; Thomas Talhelm, University of Virginia; Anna van 't Veer, Tilburg University; Michelangelo Vianello, University of Padova

## Abstract

Reproducibility is a defining feature of science. However, because of strong incentives for innovation and weak incentives for confirmation, direct replication is rarely practiced or published. The Reproducibility Project is an open, large-scale, collaborative effort to systematically examine the rate and predictors of reproducibility in psychological science. So far, 72 volunteer researchers from 41 institutions have organized to openly and transparently replicate studies published in three prominent psychological journals from 2008. Multiple methods will be used to evaluate the findings, calculate an empirical rate of replication, and investigate factors that predict reproducibility. Whatever the result, a better understanding of reproducibility will ultimately improve confidence in scientific methodology and findings.

Abstract = 109 words

Keywords = methodology, replication, reproducibility, psychological science, open

Reproducibility—the extent to which consistent results are observed when scientific studies are repeated—is one of science's defining features (Bacon, 1267/1859; Jasny, Chin, Chong, & Vignieri, 2011; Kuhn, 1962; Popper, 1934; Rosenthal, 1991),[2] and has even been described as the "demarcation criterion between science and nonscience" (Braude, 1979, p. 2). In principle, the entire body of scientific evidence could be reproduced independently by following the methods and insights gleaned by prior investigators. As such, belief in scientific evidence is not contingent on trust in its originators. For other ways of garnering knowledge, the authority and motivations of the source matter; for science, they don't.[3]

Considering its central importance, one might expect replication to be a prominent part of scientific practice. It is not (Collins, 1985; Reid, Soley, & Wimmer, 1981; Schmidt, 2009). An important reason for this is that there are strong incentives for scientists to introduce new ideas but weak incentives to confirm the validity of old ideas (Nosek, Spies, & Motyl, 2012). Innovative findings produce rewards of publication, employment, and tenure; replicating findings produce a shrug.

Taking resources away from innovation and devoting them to confirmation is a poor investment if the original findings are valid. But the costs of accepting false findings are high as well. Burgeoning research areas could expend resources fruitlessly pursuing false leads, and theories could rely on invalid empirical evidence. Wise calibration of resource investment between innovation and confirmation could take into account the reproducibility rate to maximize the rate of knowledge accumulation. How would resources be allocated if the reproducibility rate is 90%? What about 30%?

---

[2] Some distinguish between "reproducibility" and "replicability" by treating the former as a narrower case of the latter (e.g., computational sciences) or vice versa (e.g., biological sciences). We ignore the distinction.

[3] That is, they are not supposed to matter. To the extent that they do is evidence of current scientific practices relying on authority rather than evidence.

There exists very little evidence to provide reproducibility estimates for scientific fields, though some empirically informed estimates are disquieting (Ioannidis, 2005). When independent researchers tried to replicate dozens of important studies in cancer, women's health, and cardiovascular disease, only 25% confirmed the original result (Prinz et al., 2011). In a similar investigation, Begley and Ellis (2012) reported a meager 11% replication rate. In psychology, a survey of unpublished replication attempts found that about 50% replicated the original results (Hartshorne & Schachner, 2012; see also Wager et al., 2009 for neuroscience). In this paper, we introduce the Reproducibility Project: an effort to systematically estimate the reproducibility rate of psychological science as practiced currently, and to investigate factors that predict reproducibility.

**The Reproducibility Project**

Obtaining a meaningful estimate of reproducibility requires conducting replications of a sizable number of studies. However, because of existing incentive structures, it is not in an individual scientist's professional interest to conduct numerous replications. The Reproducibility Project addresses these barriers by spreading the workload over a large number of researchers. As of August 23rd, 2012, 72 volunteers from 41 institutions had joined the replication effort. Each contributor plays an important, but circumscribed, role such as contributing on a team conducting one replication study. Researchers volunteer to contribute based on interest, skills, and available resources. The project coordination, planning, materials, and execution are available publicly on the Open Science Framework (http://opensciencframework.org/). Open practices increase the accountability of the replication team and, ideally, the quality of the designs and results.

**Selecting studies for replication**. Studies eligible for replication were selected from 2008 issues of three prominent journals that differ in topical emphasis and publishing format (i.e., short report versus long-form articles): *Journal of Experimental Psychology: Learning,*

*Memory, and Cognition, Journal of Personality and Social Psychology,* and *Psychological Science*.[4] To minimize selection biases even within this restricted sample, replication teams choose from among the first 30 articles published in each journal. From the selected article, the team selects a single study (the last study unless it is unfeasible to replicate) and key finding for replication. As eligible articles are claimed, additional articles from the sampling frame are made available for selection. Not all studies can be replicated. For example, some used unique samples or specialized equipment that is unavailable, others were dependent on a specific historical event. Although feasibility constraints will reduce the generalizability of the ultimate results, they are also inevitably part and parcel of reproducibility itself.

**Conducting the replications***.* The project's replication attempts follow a standardized protocol aimed at minimizing irrelevant variation in data collection and reporting methods, and maximizing quality of replication efforts. The project attempts *direct* replications – "repetition of an experimental procedure" in order to "verify a piece of knowledge" (Schmidt, 2009, p. 92, 93). Replications must have high statistical power ($1-\beta \geq .80$ for the effect size of the original study) and use the original materials, if available. Researchers solicit feedback from the original authors on research design before data collection, particularly to identify factors that may interfere with replication. Identified threats are either remedied with revisions or coded as a potential predictor of reproducibility and written into the replication report.

**Evaluation of replication study results**. Successful replication can be defined by "vote-counting" narrowly (obtaining the same statistically significant effect as original study) or broadly (obtaining a directionally similar, but not necessarily statistically significant result), or quantitatively – for example, through meta-analytic estimates combining the original and replication study, comparisons of effect sizes, or an updated estimate of Bayesian priors. As

[4] Additional journals could be added if enough volunteers join the project.

yet, there is no single, general standard to answer "What is replication?" so we employ multiple criteria (Valentine et al., 2011).

Failures to replicate might result from several factors. The first is a simple type II error with an occurrence rate of 1-β: Some true findings will fail to replicate purely by chance. However, the overall replication rate can be compared against the average power across studies. For this reason, the project focuses on the overall reproducibility rate. Individual studies that fail to replicate are not treated as disconfirmed. Other reasons for failures to replicate include: (1) the original effect is false; (2) the actual effect is of lower size than reported originally, making it more difficult to detect; (3) the design, implementation, or analysis of either the original or replication study is flawed; or (4) the replication methodology differs from the original in ways that are critical for successful replication.[5] All reasons are important for evaluating reproducibility, but the most interesting may be the last. Identifying specific ways that replications and original studies differ, especially when replications fail, can advance theoretical understanding of the previously unconsidered conditions necessary to obtain an effect. Thus, replication is theoretically consequential.

The most important point is that a failure to replicate does not directly indicate that the original effect is false. It may also not replicate because of insufficient power, design problems, or known and unknown limiting conditions. As such, the Reproducibility Project is investigating factors such as replication power, the evaluation of the study design by the original authors, and the original study's sample and effect sizes as *predictors* of reproducibility. Identifying the contribution of these factors to reproducibility is useful because each has distinct implications for interventions to improve reproducibility.

[5] Note that the Reproducibility Project will not evaluate whether the original *interpretation* of the finding is correct. For example, if an eligible study had an apparent confound in the design, that confound would be retained in the replication attempt. Confirmation of theoretical interpretations is an independent consideration.

**Implications of the Reproducibility Project**. An estimate of the reproducibility of current psychological science will be an important first. A high reproducibility estimate might boost confidence in conventional research and peer review practices in the face of criticisms about inappropriate flexibility in design, analysis, and reporting decisions that can inflate the rate of false positives (Greenwald, 1975; John et al., 2012; Simmons et al., 2011). A low estimate might prompt reflection on the quality of standard practice, motivate further investigation of reproducibility, and ultimately lead to changes in daily practice and publishing standards (Bertamini & Munafo, 2012; LeBel & Peters, 2011).

Some may worry that discovering a low reproducibility rate will damage the image of psychology or science more generally. It is certainly possible that opponents of science will use such a result to renew their calls to reduce funding for basic research. However, we believe that there is a much worse alternative: having a low reproducibility rate, but failing to investigate and discover it. If reproducibility is lower than acceptable, then we believe it is vitally important that we know about it in order to address it. Self-critique, and the promise of self-correction, is why science is such an important part of humanity's effort to understand nature and ourselves.

## Conclusion

The Reproducibility Project uses an open methodology to test the reproducibility of psychological science. It also models procedures designed to simplify and improve reproducibility. Readers can review the discussion history of the project, examine the project design and structured protocol, retrieve replication materials from the various teams, obtain the reports or raw data from completed replications, and join the project to conduct a replication (start here: http://opensciencframework.org/project/EZcUj/). Adding to the community of volunteers will strengthen the power and impact of the project. With this open, large-scale, collaborative scientific effort, we hope to identify the factors that contribute to the reproducibility and validity of psychological science. Ultimately, such evidence – and steps toward resolution if

the evidence produces a call for action – can improve psychological science's most important asset: confidence in its methodology and findings.

**References**

Bacon, R. (1267/1859). Fr. Rogeri Bacon Opera quædam hactenus inedita. Vol. I. containing I.--Opus tertium. II.--Opus minus. III.--Compendium philosophiæ. Longman, Green, Longman and Roberts. Retrieved August 22, 2012 from: http://books.google.com/books?id=wMUKAAAAYAAJ

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*, 531-533. doi:10.1038/483531a

Bertamini, M., & Munafò, M. R. (2012). Bite-size science and its undesired side effects. *Perspectives on Psychological Science, 7*, 67-71. doi: 10.1177/1745691611429353

Braude, S. E. (1979). *ESP and psychokinesis*. A philosophical examination. Philadelphia, PA: Temple University Press.

Collins, H. M. (1985). *Changing order*. London: Sage.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20. doi:10.1037/h0076157

Hartshorne, J. K. & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience*, *6*, 8. doi: 10.3389/fncom.2012.00008

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124. doi:10.1371/journal.pmed.0020124

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again… *Science, 334,* 1225. doi: 10.1126/science.334.6060.1225

John, L., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524-532. doi: 10.1177/0956797611430953

Kuhn, T.S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*, 371-379. doi: 10.1037/a0025172

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*.

Popper, K. (1934/1992). *The Logic of Scientific Discovery*. New York, NY: Routledge.

Prinz, F., Schlange, T. & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*, 712-713. doi:10.1038/nrd3439-c1

Reid, L. N., Soley, L. C., & Wimmer, R. D. (1981). Replication in advertising research: 1977, 1978, 1979. *Journal of Advertising, 10*, 3-13. doi:10.1016/S0149-2063_03_00024-2

Rosenthal, R. (1991). Replication in behavioral research. In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 1–39). Newbury Park, CA: Sage.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90-100. doi:10.1037/a0015108

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366. doi:10.1177/0956797611417632

Valentine, J. C. Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., et al. (2011). Replication in prevention science. *Prevention Science, 12*, 103-117. doi:10.1007/s11121-011-0217-6

Wager, T. D., Lindquist, M. A., Nichols, T. E., Kober, H., & van Snellenberg, J. X. (2009). Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *Neuroimage, 45*, S210-S221. doi:10.1016/j.neuroimage.2008.10.061