





# Auditing the value of empirical audits

John G. Lynch Jr.<sup>a,1</sup>, Philip M. Fernbach<sup>a,1</sup> , and Christina Kan<sup>b</sup> 

Exact replication is impossible (1), but various approaches to “direct” and “conceptual” replication have been developed, including papers with multiple replications of a single study (2), large-scale efforts across multiple laboratories (3), and adversarial collaborations (4). Uniting these approaches is a set of best practices to ensure that published replications get the science right (5).

The O’Donnell et al. (6) “empirical audit and review” is a fraught approach likely to sow confusion. Their convenience sample of studies deters projection of findings to any meaningful universe beyond mTurk studies within papers that have cited an arbitrary target paper.

Unable to say something general about scarcity, might they add data points on the effect sizes of individual studies? For this limited aim, several of their 20 replications were neither direct nor conceptual replications (7). Pairs of PhD seminar students designed studies and collected and analyzed data without sufficient oversight. Some pairs contacted authors before data collection and were told their procedural modifications were problematic. Others did not reach out to verify methods.

Our experiences with the faulty replication of our study and its subsequent retraction show the problems with not collaborating with original authors (for details, see <https://osf.io/zyjw6/>). Our study asked in November about already-articulated holiday gift plans and an unexpected bill in November (8). Among other differences, they ran their study in April with the bill occurring at some unspecified future date. Replicators forgot to ask for gift plans. They missed data patterns that would have made methodological errors obvious to senior authors and reviewers. If the senior authors could not catch these problems, what hope could there be for the PNAS review team tasked with assessing the details of 20 studies?

Moreover, their analytical approach is flawed as it considers only the effect size from the replication, assuming it alone

reflects “truth.” (9) No original study can be assumed to be definitive, nor any imperfect replication. All are grist for later meta-analytic conclusions (1). Reporting should include a meta-analytic test of whether the original and replicate effect sizes differ. If estimates are similar, report the pooled effect size estimate and confidence bands.

We conducted these analyses on the 19 pairs of studies in their figure 1. For 14 of the 19, Q statistics indicated that original and replication might plausibly reflect samples from a common distribution. Of those 14, five had pooled effect size estimates signed like the originals with 95% CIs excluding 0. This was also true when we meta-analyzed constraint effects on the number of efficiency versus priority plans in our original November 2012 study and two replications conducted in November 2022 (see <https://osf.io/zyjw6/>).

The O’Donnell et al. replication effort had four primary weaknesses: 1) an arbitrary study universe, 2) a scope that would make it challenging if not impossible for reviewers to properly audit the studies, 3) inconsistent consultation with original authors leading to methods discrepancies, and 4) an analytical approach that ignored the original data in estimating effect sizes. The third and fourth issues might be addressed with a more careful effort, but the first and second are inherent to the form. Consequently, we view empirical audit and review to be an unpromising template for replication research.

Author affiliations: <sup>a</sup>Leeds School of Business, University of Colorado, Boulder 80309; and <sup>b</sup>School of Business, University of Connecticut, Storrs 06269

Author contributions: J.G.L., P.M.F., and C.K. wrote the paper.

The authors declare no competing interest.

Copyright © 2023 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

<sup>1</sup>To whom correspondence may be addressed. Email: philip.fernbach@colorado.edu.

Published June 20, 2023.

1. J. G. Lynch, E. T. Bradlow, J. C. Huber, D. R. Lehmann, Reflections on the replication corner: In praise of conceptual replications. *Int. J. Res. Mark.* **32**, 333–342 (2015).
2. R. A. Klein et al., Investigating variation in replicability: A “many labs” replication project. *Soc. Psychol.* **45**, 142 (2014).
3. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
4. C. J. Clark, P. E. Tetlock “Adversarial collaboration: The next science reform” in *Political Bias in Psychology: Nature, Scope, and Solutions*, C. L. Frisby, R. E. Redding, W. T. O’Donohue, S. O. Lilienfeld, Eds. (Springer, 2022).
5. B. B. McShane, J. L. Tackett, U. Böckenholt, A. Gelman, Large-scale replication projects in contemporary psychological research. *Am. Stat.* **73**, 99–105 (2019).
6. M. O’Donnell et al., Empirical audit and review and an assessment of evidentiary value in research on the psychological consequences of scarcity. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2103313118 (2023).
7. A. Shah, J. Zhao, S. Mullainathan, E. Shafir, A scarcity literature mischaracterized with an empirical audit. *PsyArXiv [Preprint]* (2018). <https://doi.org/10.31234/osf.io/gphcz> (Accessed 15 February 2023).
8. P. M. Fernbach, C. Kan, J. G. Lynch Jr., Squeezed: Coping with constraint through efficiency and prioritization. *J. Consum. Res.* **41**, 1204–1227 (2015).
9. U. Simonsohn, Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* **26**, 559–569 (2015).