REPLICATION RESEARCH

Replication Research (R2), 2026, vol 2
10.17879/replicationresearch-2026-9577
Article type: *Editorial Communication*
Discipline: *Multidisciplinary*

From Replication Rates to Replication Evidence: A Guest Editorial

Felix Holzmeister [1], Colin Camerer [2], Florian Cova [3], Anna Dreber [4], Magnus Johannesson [4]

[1]*Department of Economics, University of Innsbruck, Innsbruck, Austria*
[2]*California Institute of Technology, Pasadena, CA, USA*
[3]*University of Geneva, Geneva, Switzerland*
[4]*Department of Economics, Stockholm School of Economics, Stockholm, Sweden*

ABSTRACT.
Repeating original studies using existing data (reproducibility) or new data (replicability) is key to improving the credibility of scientific results. Here, we focus on replications, which can be broadly categorized into direct replications—testing the original hypothesis in new data using the original analysis and design—and conceptual replications—testing the original hypothesis in new data using an alternate analysis and/or design. While encouraging replications is important, increasing their visibility and impact is equally crucial. Large-scale replication projects have generated valuable insights into the overall level of replicability across fields, but they typically emphasize aggregate estimates. We argue that this focus obscures the informational value of individual replication studies. Each replication provides independent evidence for updating beliefs about the likelihood that the underlying hypothesis is true and the magnitude of the effect. Publishing and disseminating individual replications as standalone contributions can therefore enhance their role in the cumulative advancement of scientific knowledge.

KEYWORDS. replicability, knowledge accumulation, credibility, metascience.

LAY SUMMARY. This editorial argues that researchers should replicate more existing findings—that is, repeat earlier studies to see whether their results hold up. But more replications alone are not enough: these studies also need to be more visible and recognized within the scientific community. The authors highlight that individual replication studies focusing on single results are valuable in their own right. Each one provides new evidence about whether a finding is reliable and how strong its effect is. This complements large-scale replication projects, which combine many studies to give an overall picture. Treating individual replications as standalone contributions can therefore deepen our understanding beyond these broader summaries.

CITATION. Holzmeister, F., Camerer, C., Cova, F., Dreber, A., & Johannesson, M. (2026). From Replication Rates to Replication Evidence: A Guest Editorial. *Replication Research*, *2*. https://doi.org/10.17879/replicationresearch-2026-9577

Introduction

The publication of the Reproducibility Project: Psychology (RP:P) brought replications to the forefront of scientific research in the social sciences (Open Science Collaboration 2015). The project examined whether 100 key findings published in leading psychology journals could be replicated. Of the 97 studies that originally reported statistically significant results, only 36% replicated, defined as finding a statistically significant effect ($p < 0.05$) in the original direction. The RP:P was followed by several additional large-scale replication projects in the social sciences (e.g., Camerer et al. 2016; 2018; Klein et al. 2018; Cova et al., 2021; Errington et al. 2021; Davis et al. 2023; Holzmeister et al. 2025; Tyner et al. 2026). Taken together, these efforts suggest a replication rate of approximately 50%, both in terms of the share of replications yielding statistically significant effects in the original direction and in terms of the magnitude of replicated effect sizes relative to the originals.[1]

While these large-scale replication efforts have had a substantial impact in raising awareness of the limited credibility of published findings and the importance of replications, the results of individual replication studies often receive limited attention and are not effectively disseminated once aggregated into large-scale projects. In what follows, we discuss replications of quantitative studies in more detail, starting with defining key concepts and outlining the characteristics of high-quality replication studies. We then consider the impact of replications and how their contribution to the accumulation of knowledge can be improved. In particular, we argue that largely treating individual replications as inputs into aggregate replicability statistics underutilizes the informational value of replication studies and precludes effective belief updating and efficient knowledge accumulation. We conclude with several suggestions for future practice.

Key concepts

There has been considerable confusion in the literature regarding the meaning of terms such as reproducibility and replicability. In recent years, a consensus has emerged defining reproducibility as the ability to regenerate original results using the same data, and replicability as the ability to regenerate findings in new data. Under these definitions, the RP:P assessed replicability rather than reproducibility. Following the terminology in Dreber & Johannesson (2025a), reproducibility can be divided into computational, recreate, and robustness reproducibility, while replicability can be divided into direct and conceptual replicability. These distinctions are not only conceptual but also have implications for how replication evidence is generated, interpreted, and ultimately disseminated—an issue we return to below.

Computational reproducibility has a long tradition in the social sciences (e.g., Dewald et al. 1986) and refers to the ability to reproduce original results using the authors' data and code. While computational reproducibility has been argued to constitute a minimum requirement

[1] For a regularly updated overview of large-scale replication projects—including both completed and ongoing efforts across the social sciences and humanities—see the FORRT Replication Hub, which offers a useful entry point for tracking developments in large-scale replication research.

for credible and trustworthy empirical research (e.g., Stark 2018), evidence suggests that it has been disappointingly low in many settings (Chang & Li 2017; Gertler et al. 2018; Perignon et al. 2024). However, it remains unclear whether and to what extent failures of computational reproducibility introduce systematic bias in reported effect sizes or instead resemble random measurement error. Somewhat surprisingly, the question of systematic bias has received little attention in this literature. The increasing use of data editors at journals—who verify that submitted data and code reproduce reported results prior to publication—can be expected to substantially improve computational reproducibility (e.g., Vilhuber & Cavanagh 2025), even though some coding and data errors may still go undetected.

Recreate reproducibility refers to attempts to recreate published results without access to the original code and/or dataset (while using the same data sources and analytical procedures as described in the paper as closely as possible). Recent evidence suggests that such efforts tend to yield weaker support for the tested hypotheses (Delios et al. 2022; Black et al. 2024).

Robustness reproducibility, or simply robustness, examines how sensitive published findings are to alternative analytical choices applied to the same data. Several recent large-scale robustness studies point to the limited robustness of many published findings (Campbell et al. 2024; Aczel et al. 2026; Brodeur et al. 2024). Because no new data collection is required, robustness analyses are generally less resource-intensive than replication studies and may be more prevalent nowadays. However, their ability to inform about true underlying effect sizes is limited, as any biases in the original sample carry over to the robustness tests. In addition, assessing the quality of robustness analyses is arguably more challenging than evaluating replication studies.

Replication studies can be categorized into direct and conceptual replications. *Direct replications* aim to implement the same design and analysis in a new sample, whereas *conceptual replications* test the same hypothesis using alternative designs and/or analytical approaches in new data. Dreber & Johannesson (2025a) further distinguish replications based on whether the new data come from the same, a similar, or a different population. Replications drawing on the same population are rare, as they require collecting independent samples from the same population.

In practice, the boundary between direct and conceptual replications is often blurred, as even minor design changes can make classification ambiguous. Nevertheless, the distinction is important, and conceptual replications differ in a fundamental way, as they are less prone to inheriting biases embedded in the original study's design or analysis. For example, if an original analytical approach systematically overestimates effect sizes, direct replications using the same analysis will inherit this bias, whereas conceptual replications that implement an alternative analysis may not. Since most large-scale replication projects rely on direct replications, their estimates of effect sizes may be systematically upward-biased. At the same time, many studies that effectively function as conceptual replications are not labeled as such in the published literature, possibly because framing a study as a replication is perceived to reduce publication prospects.

A third category of replication studies, which has received little attention yet, is *multiverse replications*. These involve conducting a multiverse analysis (Patel et al. 2015, Steegen et al. 2016, Simonsohn et al. 2020) in an independent sample, systematically exploring a multitude of reasonable analytical choices. Multiverse replications combine elements of replication and robustness tests: they mitigate biases tied to specific analytical choices and reduce the impact of researchers' degrees of freedom. While still rare, they represent a promising direction, particularly if extended to variations in both research designs and analyses (Huber et al. 2023).

Why do studies fail to replicate?

Evidence from large-scale replication projects shows that many studies reporting statistically significant findings fail to replicate. Within the conventional null hypothesis testing model, several mechanisms can account for such failures (Ioannidis 2005). First, testing hypotheses with low prior probabilities increases the likelihood that statistically significant findings are false positives (Dreber et al. 2015; Johnson et al. 2017). Second, low statistical power inflates the false positive rate (Button et al. 2013). The bias arises from the selection of results based on statistical significance: in low-powered studies, estimated effect sizes must exceed the true effect size to reach significance, leading to systematically inflated estimates.

Beyond the conventional framework, additional factors such as heterogeneity in effect sizes and selective reporting of results can further contribute to replication failures. Analytical and design heterogeneity imply that different yet justifiable analytical choices and research designs can yield systematically different effect-size estimates. As a result, analyses or designs that produce larger effects are more likely to achieve nominal statistical significance and be reported, generating false positives (Holzmeister et al. 2024). Importantly, such bias can occur even in preregistered studies that rely on a single analysis and design.

Selective and opportunistic reporting of results—often referred to as *p*-hacking—constitutes another important source of bias (Simmons et al. 2011; Gelman & Loken 2014; Brodeur et al. 2016, 2020). Researchers may, intentionally or unintentionally, report only a subset of conducted analyses, typically those yielding larger effect sizes and/or smaller *p*-values, thereby increasing the likelihood of publication. This selective process further amplifies distortions in the published evidence base and contributes to the low rate of successful replications.

The above reasons for replication failures are based on the original findings being false positives. But it should also be noted that true original claims may fail to replicate due to insufficient replication power, a point we return to below.

How should we measure replicability?

There is no consensus on how best to measure replicability, and a range of indicators has been proposed in the literature (Heyard et al. 2025). Still, two indicators are the most widely used for original results reported as statistically significant. The first is the *statistical significance indicator*, which defines a successful replication as one that yields a *p*-value <

0.05 and an effect in the same direction as the original study. The second is the *relative effect size indicator*, which measures replicability as the ratio of the replication effect size to the original effect size.[2]

While these two replication indicators are closely related at the aggregate level, they capture slightly different aspects of replicability. The statistical significance indicator is sensitive to the replication study's statistical power, whereas the relative effect size indicator also reflects the inflation of effect sizes in original studies, accounting for both false positives and exaggerated true positives.

With high replication power, the two measures are mechanically linked: the relative effect size indicator approximates the product of the share of statistically significant replications and the mean relative effect size among those studies that successfully replicated according to the significance indicator. For example, if 50% of studies replicate according to the significance indicator and the mean relative effect size among these studies is 80%, the mean relative effect size will be 40% (0.5 × 0.8; see Dreber & Johannesson 2025b).

As a measure of the aggregate replication rate across a set of studies, the average relative effect size is preferable, as it is not directly affected by replication power and captures the inflation of effect sizes not only for false positives but also for true positives. At the same time, the two indicators address complementary questions: (i) whether the original hypothesis receives statistical support in new data, and (ii) how large the replication effect size is relative to the original estimate. Alternative measures of replicability, such as, e.g., the small-telescopes approach (Simonsohn, 2015) or prediction intervals (Patil et al. 2016), answer complementary questions but come with their own caveats and limitations (for a discussion, see, e.g., Camerer et al. 2018 and Holzmeister et al. 2025). In most applications, the statistical significance indicator and the relative effect size indicator are often sufficient to draw reliable conclusions about replicability, especially in the context of meta-studies that aggregate results from several replication attempts. For individual replications, it is further worthwhile to estimate how prior beliefs about the tested hypotheses are affected by the replication outcome, which captures an additional and important dimension of replicability (see below for details).

What do we learn from replications?

Replications contribute to the accumulation of knowledge by providing new evidence on whether a tested hypothesis is likely true or false and on the magnitude of the true underlying effect. From a Bayesian perspective, this implies updating prior beliefs about the probability of the hypothesis being genuinely true. A successful replication increases this probability, whereas a failed replication decreases it.

[2] Note that these two indicators apply to original results reported as statistically significant. The replication project led by Cova et al. (2021) also included studies whose main conclusion was the *absence* of a statistically significant effect. For further discussion of how to assess replications of null findings, see Pawel et al. (2024).

Using data from the RP:P, Dreber et al. (2015) estimate that a successful replication—defined in terms of the significance indicator—increases the median probability that a hypothesis is true from 56% to 98%, while a failed replication reduces it to 6%. While these estimates are specific to the RP:P, they illustrate that replications can substantially shift normative beliefs about the likelihood that a hypothesis is true and about the validity of scientific claims.

At the same time, quantifying belief updates for individual replication studies is not straightforward, as it requires information or assumptions about the prior probability before conducting the replication. Moreover, realizing and fully harnessing the informational value of replications critically depends on their effective communication and integration into the scholarly literature. This is why it is important that the detailed methods and results of each individual replication study included in large-scale replication projects are made accessible to researchers (see below).

Carrying out high-quality replications

How can we distinguish between high- and low-quality replications? A first key criterion is whether the replication study was preregistered with a pre-analysis plan (or conducted as a Registered Report), and whether the preregistration was sufficiently detailed and adhered to in the analysis (Nosek et al. 2018; Chambers & Tzavella 2022; Dreber & Johannesson 2025b). Ideally, all replication studies should follow either of these two approaches. For pre-registration to increase the credibility of published findings, it is crucial that the pre-analysis plan is detailed and strictly followed; see Dreber & Johannesson (2025b) for more details.

A second critical factor is statistical power. Importantly, even true-positive findings tend to have inflated effect sizes due to insufficient statistical power in the original studies, a fact that ought to be accounted for when designing replications. Some of the early large-scale replication projects, such as the RP:P (Open Science Collaboration 2015) and the Experimental Economics Replication Project (EERP; Camerer et al. 2016), did not fully incorporate this consideration and were therefore underpowered. The insufficient replication power likely contributed to the relatively low replication rate of 36% reported in the RP:P, which may underestimate the true share of genuine effects. Later efforts, such as the Social Sciences Replication Project (SSRP; Camerer et al. 2018) or the MTurk Replication Project (MTRP; Holzmeister et al. 2025), addressed this issue by substantially increasing replication power.

A useful way to think about replication power is as the fraction of the original effect size that the replication is designed to detect with high probability. As a benchmark informed by meta-analytic evidence from previous large-scale projects, we recommend aiming for 90% power to detect half to two-thirds of the original effect size. For relatively low-powered original findings, achieving high replication power typically requires substantially larger sample sizes than those in the underlying study. For example, if the original result is only "marginally significant" (with a *p*-value close to the 5% threshold), designing a replication with 90% power to detect one-half [two-thirds] of the original effect size requires a replication sample size that is roughly ten times [six times] larger than the original.

The quality of direct replications is generally easier to assess than that of conceptual replications, as they can be evaluated—at least partially—based on how closely they follow the original design and analysis. When original authors make their materials and code available, a direct replication can closely mirror the original study. In such cases, beyond preregistration and adequate statistical power, the key remaining factor for direct replications is the population from which the replication sample is drawn. Ideally, this population should be as similar as possible to the original.[3] At the same time, however, systematically examining how effect sizes vary across populations—as, e.g., in the Many Lab studies (Klein et al. 2014, 2018)—provides valuable additional insights into the nature of the phenomenon under investigation and into potential moderating effects that contribute to effect-size heterogeneity.

Several large-scale replication projects have further enhanced quality by sharing detailed replication plans with the original authors for feedback and approval prior to data collection, and by publicly posting (i.e., preregistering) these plans in advance. This practice can be valuable, although it depends on the willingness of original authors to engage.

Assessing the quality of conceptual replications is generally more challenging, as it requires evaluating whether alternative designs and analytical approaches provide appropriate tests of the original hypothesis. This challenge strengthens the case for multiverse replications, which systematically explore a range of justifiable analytical choices and, potentially, design variations. More broadly, the "status" of conceptual replications should be elevated, as they can help prevent biases embedded in the original analysis from persistently biasing tests of specific hypotheses. Registered Reports publications are particularly well-suited for this purpose, as they allow for critical peer review of designs, analyses, and populations prior to data collection.

The impact of replications

As discussed above, replications can—if conducted rigorously—substantially shift beliefs about whether a hypothesis is true and help distinguish likely true positives from likely false positives. However, this potential is only realized if individual replication outcomes are visible, interpretable, and accessible as independent pieces of evidence.

In this respect, the broad message from large-scale replication projects—that replicability in the social sciences and beyond is limited—has been widely communicated (Baker 2016). Terms such as "replication crises" and "reproducibility crises" gained prominence following the publication of the RP:P, which has been cited extensively. Several subsequent large-scale replication projects have likewise received considerable attention.

In contrast, the results of individual replications synthesized within these large-scale projects are often not treated as standalone scholarly contributions, but instead remain embedded within project-level summaries. One way to indirectly assess this is by examining citation

[3] What counts as 'similar' will depend on the hypotheses tested and phenomena investigated. For example, whether two samples share the same religious beliefs might be more relevant for a study on moral attitudes than for a study on memory skills.

patterns of the original studies. If replication outcomes are incorporated into scientific beliefs, successfully replicated studies should receive more citations, while studies that fail to replicate should receive fewer. Empirical evidence on this question is mixed. Schafmeister (2021) and Serra-Garcia & Gneezy (2021) find no effect of replication outcomes on citations, whereas Clark et al. (2023) document a 14% decline in citations following failed replications. Overall, these results suggest that belief updating in response to replication evidence is limited.

A key implication is that initial false positive results may impose substantial and persistent costs. Early published findings can benefit from a "first-mover advantage," shaping beliefs, citations, and research agendas even when later contradicted by replication evidence. This persistence is consistent with "belief perseverance," as documented in cognitive psychology (Ross et al. 1975; Anderson et al. 1980). Crucially, the limited visibility of individual replication results can reinforce this dynamic by constraining their ability to counter influential yet potentially spurious findings. This underscores the need for more effective dissemination of replication results—especially at the level of individual studies—as well as for improving the quality of original studies to reduce the prevalence of false positives in the first place.

Improving the accumulation of knowledge and moving forward

Improving the dissemination of replication evidence can be approached in several ways. One important step is to complement the focus on meta-analytic findings from large-scale replication projects with greater attention to individual replication outcomes. Publishing replications as standalone contributions is a central step in this direction, as it allows individual replication results to be evaluated through peer review and integrated into the literature as independent evidence—a role that dedicated outlets such as *Replication Research* (R2) are designed to fulfill.

Raising the status of replications within the profession is equally important, as it helps improve incentives to conduct and publish replication work. In addition, systematically collecting information on replication outcomes in centralized, easily accessible databases—such as the FORRT Library of Reproduction and Replication Attempts (FLoRA)—facilitates identifying existing replications and conducting meta-research on them (Wallrich et al. 2026). For direct replications in particular, linking replication evidence to original publications—for example, through references in the original study's online version or citation alerts—could further enhance visibility and facilitate belief updating. The reference management tool *Zotero provides* an integrated retraction-detection feature through its partnership with *Retraction Watch*. Tondlekar et al. (2026) recently introduced a similar functionality for replications by linking replication information in FLoRA to references in Zotero (see https://forrt.org/flora_zotero). This is a potentially valuable tool for improving the visibility and use of replication evidence going forward.

At the same time, reducing the prevalence of false positives in original research remains essential. This can be achieved by increasing statistical power, adopting more stringent significance thresholds (e.g., $p < 0.005$; Benjamin et al. 2018), and prioritizing transparency

through preregistration and Registered Reports to counteract questionable research practices and publication bias (Nosek & Lakens 2014; Nosek et al. 2018). More broadly, there is a need for larger and more systematic original studies that explicitly account for heterogeneity by varying designs, analyses, and populations (Holzmeister et al. 2024), offering a promising path toward more informative and reliable scientific knowledge.

We are also likely to see an increasing role for artificial intelligence-based methods in supporting research on reproducibility and replicability. Such tools have already proven useful for coding, statistical analysis, language processing, and broader scientific workflows, and their capabilities are improving rapidly. AI methods appear particularly well-suited for assessing computational reproducibility and facilitating robustness analyses at scale. They may also play an important role in replication work based on retrospective data by systematically applying comparable analyses across alternative datasets. Their role in replication studies involving prospective data collection, such as experiments, is less clear, but they could, for example, contribute by assessing prior distributions to estimate posterior probabilities that hypotheses are true.

Social Impact and Responsibility

Replication research plays a central role in the accumulation of scientific knowledge by systematically updating and correcting previously published findings. This has important positive social implications, as it improves the reliability of the evidence base used by researchers, policymakers, and the broader public. By providing more accurate information about the likelihood that specific hypotheses are true and about the magnitude of underlying effects, replication work contributes to better-informed decision-making.

At the same time, replication findings can be misinterpreted. Failed replications should not be viewed as definitive evidence against an effect, and successful replications do not imply proof of the hypothesis nor generalizability. Clear communication, emphasizing uncertainty and context, is therefore essential. Replications also affect perceptions of scientific credibility. While documenting limited replicability may challenge confidence in specific findings, it strengthens trust by demonstrating transparency and self-correction. Improving the reliability and interpretation of scientific evidence helps reduce risks associated with acting on false positives or exaggerated effects, which can have broad societal consequences.

Declarations

Author Contributions

- ✦ Writing, original draft: F.H., M.J.
- ✦ Writing, review and editing: F.H., C.C., F.C., A.D., M.J.

Acknowledgments

We thank Flavio Azevedo, Lukas Röseler, and Lukas Wallrich for the invitation to write this Guest Editorial and for helpful feedback.

Funding

For financial support, we thank the Jan Wallander and Tom Hedelius Foundation (grants P23-0098 and P25-0210 to A.D.) and the Knut and Alice Wallenberg Foundation (grant KAW 2023.0363 to A.D.).

Potential Conflicts of Interest

The authors declare no conflicts of interest.

Declaration of AI use

No generative large-language models were used in the writing of this manuscript. Grammarly was used to improve the manuscript's grammar, spelling, and language clarity. The authors reviewed and approved all suggestions made by Grammarly and take full responsibility for the final text.

Author Contact

Correspondence should be addressed to Felix Holzmeister, email: Felix.Holzmeister@uibk.ac.at.

References

Aczel, B., Szaszi, B., Clelland, H. T., Kovacs, M., Holzmeister, F., Van Ravenzwaaij, D., ... Geraldes, D. (2026). Investigating the analytical robustness of the social and behavioural sciences. *Nature*, *652*(8108), 135-142. https://doi.org/10.1038/s41586-025-09844-9

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, *39*(6), 1037-1049. https://doi.org/10.1037/h0077720

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452-454. https://doi.org/10.1038/533452a

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6-10. https://doi.org/10.1038/s41562-017-0189-z

Black, B., Desai, H., Litvak, K., Yoo, W., & Yu, J. J. (2024). The SEC's short-sale experiment: Evidence on causal channels and reassessment of indirect effects. *Management Science*, *70*(8), 4953-5625. https://doi.org/10.1287/mnsc.2023.4918

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review*, *110*(11), 3634-3660. https://doi.org/10.1257/aer.20190687

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star Wars: The empirics strike back. *American Economic Journal: Applied*, *8*(1), 1-32. https://doi.org/10.1257/app.20150044

Brodeur, A., Mikola, D., & Cook, N. (2024). Mass reproducibility and replicability: A new hope (IZA Discussion Paper No. 16912). IZA Institute of Labor Economics. https://ssrn.com/abstract=4790780

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433-1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H.(2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637-644. https://doi.org/10.1038/s41562-018-0399-z

Campbell, D., Brodeur, A., Dreber, A., Johannesson, M., Kopecky, J., Lusher, L., & Tsoy, N. (2024). The robustness reproducibility of the American Economic Review. I4R Discussion Paper Series (No. 124). https://www.econstor.eu/bitstream/10419/295222/1/I4R-DP124.pdf

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, *6*(1), 29-42. https://doi.org/10.1038/s41562-021-01193-7

Chang, A. C., & Li, P. (2017). A preanalysis plan to replicate sixty economics research papers that worked half of the time. *American Economic Review*, *107*(5), 60-64. https://doi.org/10.1257/aer.p20171034

Clark, C. J., Connor, P., & Isch, C. (2023). Failing to replicate predicts citation declines in psychology. *Proceedings of the National Academy of Sciences*, *120*(29), e2304862120. https://doi.org/10.1073/pnas.2304862120

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye, N., van Dongen, N., Dranseika, V., Earp, B. D., Gaitán Torres, A., Hannikainen, I., Hernández-Conde, J. V., … Zhou, X. (2021). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, *12*(1), 9-44. https://doi.org/10.1007/s13164-018-0400-9

Davis, A. M., Flicker, B., Hyndman, K., Katok, E., Keppler, S., Leider. S., et al. (2023). A replication study of operations management experiments in Management Science. *Management Science*, *69*, 4977-4991. https://doi.org/10.1287/mnsc.2023.4866

Delios, A., Clemente, E. G., Wu, T., Tan, H., Wang, Y., Gordon, M., … & Uhlmann, E. L. (2022). Examining the generalizability of research findings from archival data. *Proceedings of the National Academy of Sciences*, *119*(30), e2120377119. https://doi.org/10.1073/pnas.2120377119

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in empirical economics: The Journal of Money, Credit and Banking project. *The American Economic Review*, *76*(4), 587–603.

Dreber, A., & Johannesson, M. (2025a). A framework for evaluating reproducibility and replicability in economics. *Economic Inquiry*, *63*(2), 338-356. https://doi.org/10.1111/ecin.13244

Dreber, A., & Johannesson, M. (2025b). *The credibility gap: Evaluating and improving empirical research in the social sciences*. Routledge.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343-15347. https://doi.org/10.1073/pnas.1516179112

Errington, T., Mathur, M., Soderberg, C.K., Denis, A., Iorns, E., & Nosek, B.A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*, e71601. https://elifesciences.org/articles/71601

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460-465. https://doi.org/10.1511/2014.111.460

Gertler, P., Galiani, S., & Romero, M. (2018). How to make replication the norm. *Nature*, *544*(7693), 417-419. https://doi.org/10.1038/d41586-018-02108-9

Heyard, R., Pawel, S., Frese, J., Voelkl, B., Würbel, H., McCann, S., Held, L., Wever, K. E., Hartmann, H., Townsin, L., & Zellers, S. (2025). A scoping review on metrics to quantify reproducibility: A multitude of questions leads to a multitude of metrics. *Royal Society Open Science*, 12(7), Article 242076.https://doi.org/10.1098/rsos.242076

Holzmeister, F., Johannesson, M., Böhm, R., Dreber, A., Huber, J., & Kirchler M. (2024). Heterogeneity in effect size estimates. *Proceedings of the National Academy of Sciences*, *121*(32), e2403490121. https://doi.org/10.1073/pnas.2403490121

Holzmeister, F., Johannesson, M., Camerer, C. F., Chen, Y., Ho, T.-H., Hoogeveen, S., Huber, J., Imai, N., Imai, T., Jin, L., Kirchler, M., Ly, A., Mandl, B., Manfredi, D., Nave, G., Nosek, B. A., Pfeiffer, T., Sarafoglou, A., Schwaiger, R., … Dreber, A. (2025). Examining the replicability of online experiments selected by a decision market. *Nature Human Behaviour*, *9*(2), 316-330. https://doi.org/10.1038/s41562-024-02062-9

Huber, C., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Weitzel, U., Abellán, M., Adayeva, X., Ay, F. C., Barron, K., Berry, Z., Bönte, W., Brütt, K., Bulutay, M., Campos-Mercade, P., Cardella, E., Claassen, M. A., Cornelissen, G., Dawson, I. G. J., … Holzmeister, F. (2023). Competition and moral behavior: A meta-analysis of forty-five crowd-sourced experimental designs. *Proceedings of the National Academy of Sciences*, *120*(23), e2215572120. https://doi.org/10.1073/pnas.2215572120

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.1004085

Johnson, V. E., Payne, R.D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association*, *112*(517), 1-10. https://doi.org/10.1080/01621459.2016.1240079

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A.(2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490. https://doi.org/10.1177/2515245918810225

Nosek, B. A., Ebersole, C.R., DeHaven, A.C., & Mellor, D.T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*(9), 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should we expect when we replicate? A statistical view of replicability in psychological science. *Perspectives of Psychological Science*, *11*(4), 539-544. https://doi.org/10.1177/1745691616646366

Pawel, S., Heyard, R., Micheloud, C., & Held, L. (2024). Replication of null results: Absence of evidence or evidence of absence? *eLlife*, *12*, RP92311. https://doi.org/10.7554/eLife.92311.3

Pérignon, C., Akmansoy, O., Hurlin, C., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Menkveld, A. J., Razen, M., & Weitzel, U. (2024). Computational reproducibility in finance: Evidence from 1,000 tests. *The Review of Financial Studies*, *37*(11), 3558-3593. https://doi.org/10.1093/rfs/hhae029

Ross, L., Lepper, M.R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, *32*(5), 880-892. https://doi.org/10.1037/0022-3514.32.5.880

Schafmeister, F. (2021). The effect of replications on citation patterns: Evidence from a large-scale reproducibility project. *Psychological Science*, *32*(10), 1537-1548. https://doi.org/10.1177/09567976211005767

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), eabd1705. https://doi.org/10.1126/sciadv.abd1705

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*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559-569. https://doi.org/10.1177/0956797614567341

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Stark, P. B. (2018). Before reproducibility must come preproducibility. *Nature*, *557*(7706), 613–614. https://doi.org/10.1038/d41586-018-05256-0

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712. https://doi.org/10.1177/1745691616658637

Tondlekar, R., Wallrich, L., Weinerova, J., Fouilloux, A., Baldoni, C., Meier, M., Flores Kanter, P. E., Paiva Trajano, I., Vaidis, D. C., Müller, M., Arriaga Ferreira, P., Coulson, H., Röseler, L., & FORRT. (2026). Zotero Replication Checker (Version 0.1.12) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.18671300

Tyner, A. H., Abatayo, A.L., Daley, M., Field, S., Fox, N., Haber, N.A., et al. (2026). Investigating the replicability of the social and behavioural sciences. *Nature*, *652*(8108), 143-150. https://www.nature.com/articles/s41586-025-10078-y

Vilhuber, L., & Cavanagh, J. (2025). Report of the AEA Data Editor. *AEA Papers and Proceedings*, 113, 850–863. https://doi.org/10.1257/pandp.115.944

Wallrich, L., Röseler, L., Hartmann, H., Ashcroft-Jones, S., Doetsch, C., Kaiser, L., Schüller, S. M., Aldoh, A., Behbood, H., Elsherif, M. M., Klett, N., Krapp, J., Liu, M., Pavlović, Z., Pennington, C. R., Schütz, A., Seida, C., Siziva, K., Skvortsova, A., … Azevedo, F. (2026). FORRT Library of Replication Attempts (FLoRA) [Data set]. OSF. https://doi.org/10.17605/OSF.IO/9R62X

| REPLICATION RESEARCH R2 | 2026

doi.org/10.17879/replicationresearch-2026-9577

Replication Research (R2, ISSN: 3052-5977) is part of the Framework for Open and Reproducible Research Training (FORRT; forrt.org) and the Münster Center for Open Science (MüCOS; uni-muenster.de/MueCOS)

MüCOS | FORRT | R2