RESEARCH · RESEARCH INTEGRITY

The Replication Crisis: Causes, Evidence and Reforms

What the replication crisis is, the large replication projects that revealed it, its causes in statistics, publication and incentives, the debate over how serious it is, and the reforms it has prompted.

ModQu research article · 2,735 words · 25 references (APA 7th edition)

A single study is meant to be a step toward reliable knowledge, not the last word. The assumption behind publishing results is that other researchers can repeat a study and obtain similar findings. In the 2010s, a series of large-scale attempts to do exactly that produced unsettling results: in several fields, a large share of published findings could not be reproduced. The episode became known as the **replication crisis**. This article explains what happened, why it happened, how serious it is, and what has changed as a result.

Reproducibility and replicability

Two terms are often used loosely. The National Academies of Sciences, Engineering, and Medicine (2019) proposed a distinction that is now widely followed. **Reproducibility** means obtaining the same results from the same data using the same methods and code. **Replicability** means obtaining consistent results when a study is repeated with new data. A study can fail on either count: its published numbers may not follow from its own data, or a new study may not find the same effect. The replication crisis concerns mainly the second.

Replications also differ in how closely they follow the original. A **direct** (or close) replication repeats the original procedures as faithfully as possible. A **conceptual** replication tests the same hypothesis with different methods or measures. Both have value, but only direct replications can show whether a particular finding holds up, because a failed conceptual replication may simply reflect the changes in method.

How the crisis came to light

Concern had been growing for years. Ioannidis (2005) argued on statistical grounds that, given small studies, small effects, flexible analyses and many teams testing many hypotheses, a majority of published positive findings in some fields were likely to be false. His argument was theoretical, but it drew attention to how the conditions of research shape the reliability of its results.

Two events in 2011 sharpened the concern in psychology. Bem (2011) published nine experiments in a leading journal reporting evidence that people could sense future events. The studies followed standard methods and passed peer review, yet few researchers believed the conclusion. If standard methods could produce evidence for something so implausible, the methods themselves were in question. In the same year, Simmons et al. (2011) showed that common, undisclosed

choices in collecting and analysing data could make almost any hypothesis appear to be supported. In one demonstration, using flexible but individually defensible choices, they produced a statistically significant result showing that listening to a particular song made participants younger. Their simulations showed that combining a few such choices could raise the chance of a false positive from the nominal 5 per cent to more than 60 per cent.

Pashler and Wagenmakers (2012) described the mood in psychology as a "crisis of confidence", and researchers began to test the question directly.

The large replication projects

The most influential test was the Reproducibility Project: Psychology. The Open Science Collaboration (2015) repeated 100 studies published in three leading psychology journals, working where possible with the original authors and using larger samples. Of the original studies, 97 per cent had reported statistically significant results. Only 36 per cent of the replications did. On average, the effects found in the replications were about half the size of the original effects.

Other fields produced similar warnings. In preclinical cancer research, Begley and Ellis (2012) reported that scientists at a biotechnology company had been able to confirm the findings of only 6 of 53 landmark studies. A year earlier, Prinz et al. (2011) had reported that in-house work at a pharmaceutical company matched published findings in only about a fifth to a quarter of the drug-target projects reviewed. Both reports described internal company checks rather than published replications, and neither identified the studies concerned, but they suggested that the problem was not confined to psychology.

In economics, Camerer et al. (2016) replicated 18 laboratory experiments published in two leading journals and found a significant effect in the same direction as the original in 11, or 61 per cent. The rate was higher than in psychology, but the replicated effects were again smaller than the originals on average.

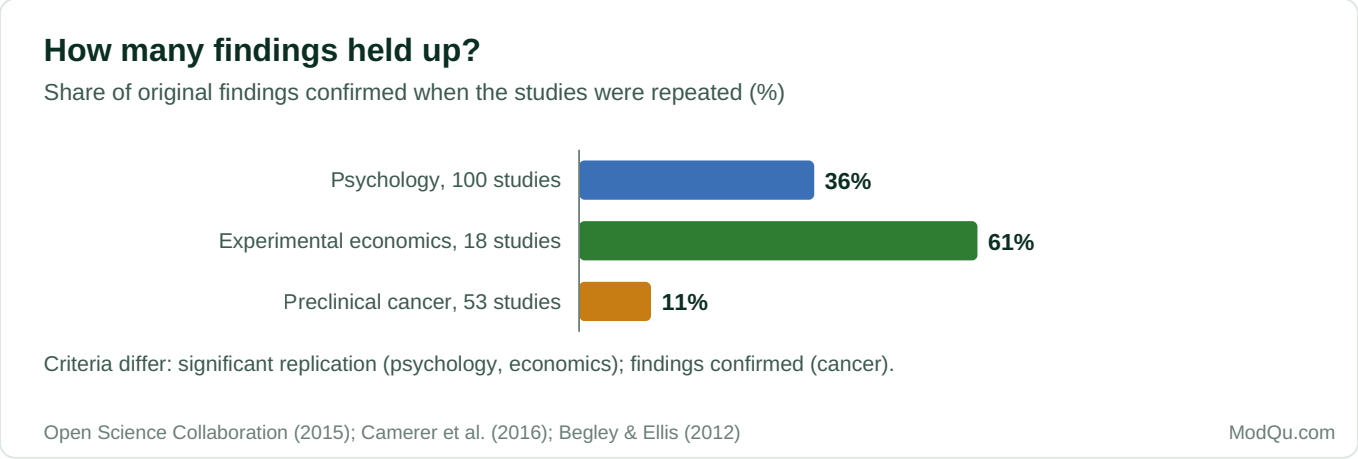


Figure 1. Share of findings confirmed in three replication efforts. Criteria differ between projects, so the figures are not directly comparable. Data from the Open Science Collaboration (2015), Camerer et al. (2016) and Begley and Ellis (2012); chart by ModQu.

The problem was also visible in what researchers reported about their own experience. In a survey of more than 1,500 scientists conducted by *Nature*, more than 70 per cent said they had tried and failed to reproduce another scientist's experiment, and a majority said there was a significant crisis of reproducibility (Baker, 2016).

Why so many findings failed to replicate

No single cause explains the crisis. Several features of how research was done and rewarded combined to produce a literature in which many published effects were exaggerated or false.

Low statistical power

A study's **power** is the probability that it will detect an effect of a given size if the effect is real. Cohen (1962) reviewed studies in a leading psychology journal and found that their average power to detect a medium-sized effect was below 50 per cent. Button et al. (2013) estimated that the median power of studies in neuroscience was around 20 per cent. Low power has two consequences that are often missed. First, many real effects go undetected. Second, the effects that do reach significance in small studies tend to be overestimates, because only the studies that happened to observe unusually large effects cross the threshold. When such studies are replicated with larger samples, the effects shrink.

Flexible analysis

Every study involves many decisions: which participants to exclude, which outcomes to report, which control variables to include, when to stop collecting data. Gelman and Loken (2014) described this as a **garden of forking paths**: even a researcher with no intention to mislead, making reasonable choices in response to the data, can arrive at a significant result that would not have appeared under a different but equally reasonable path. Simmons et al. (2011) showed how quickly such flexibility inflates false positives.

Questionable research practices

Some practices go further. John et al. (2012) surveyed about 2,000 psychologists and found that substantial shares admitted to practices such as failing to report all of a study's measures, deciding whether to collect more data after looking at the results, and reporting only the studies that "worked". Kerr (1998) described **HARKing**, hypothesising after the results are known, in which a hypothesis suggested by the data is presented as if it had been predicted in advance. These practices are not fraud, and many researchers did not regard them as wrong, but they make published results look stronger than the evidence warrants.

Publication bias

Journals have long favoured positive, novel results. Rosenthal (1979) described the **file drawer problem**: studies with null results are less likely to be published, so the published record overstates effects. Fanelli (2010) found that more than 90 per cent of papers in psychology and psychiatry reported support for their hypotheses, a higher share than in the physical sciences. Franco et al. (2014) followed a set of social science experiments from start to finish and found that strong results were far more likely to be published than null results, most of which were never written up.

Incentives

Behind these practices lie the incentives of academic careers. Hiring, promotion and funding reward the number of publications and their prominence, and journals reward novelty. Smaldino and McElreath (2016) modelled how these incentives operate across generations of researchers. Labs that produce more publications, whatever their methods, are more likely to be imitated and to train the next generation, so weak methods can spread even when no individual intends to cut corners. They noted that the statistical power of studies in the behavioural sciences had not improved over several decades despite repeated warnings.

Misunderstanding of p-values

Many problems were made worse by treating a p-value below 0.05 as proof that an effect is real. The American Statistical Association issued a statement on p-values to correct common misunderstandings, stressing that a p-value does not measure the probability that a hypothesis is true or the size or importance of an effect, and that scientific conclusions should not rest on whether a p-value passes a threshold (Wasserstein & Lazar, 2016).

What a failed replication can and cannot show

Interpreting replication results requires some care, because chance affects replications as much as originals. Consider an illustrative case. Suppose a real effect exists and a replication is designed with 80 per cent power to detect it. One time in five, the replication will fail to reach significance even though the effect is real. If the original study overestimated the effect, as small significant studies tend to do, and the replication was powered to detect the inflated size, the true power is lower still and failures become more likely.

The opposite error is also possible. If the original finding was a false positive, a replication will still reach significance about one time in twenty by chance. A single replication therefore shifts the balance of evidence but rarely settles a question on its own.

For this reason, the large projects did not rely only on whether each replication was significant. They also compared effect sizes, checked whether the original estimate fell within the replication's confidence interval, and in some cases combined the original and replication data. Each criterion gives a somewhat different rate, which is one reason why published replication rates vary. What gave the projects their weight was not any single number but the consistent pattern: across criteria, replicated effects were smaller and less often significant than the originals.

There is also a difference between failing to replicate and failing to reproduce. When independent teams have re-run the original analyses using the authors' own data, they have sometimes found errors in coding, data handling or reporting. Such failures of reproducibility are often easier to fix than failures of replicability, but they show why sharing data and code matters.

How serious is the crisis?

Not everyone accepted the most pessimistic reading. Gilbert et al. (2016) argued that many replications in the Reproducibility Project had differed from the originals in important ways, such as using different populations, and that the results were consistent with a reasonably high rate of reproducibility once such differences and chance were taken into account. Their critique prompted a vigorous reply from the original team, and the exchange showed that interpreting a failed replication is itself difficult.

Several points are now widely accepted. A single failed replication does not prove that an original finding was false, just as a single successful study did not prove it true. Replication rates depend on the field, the type of study and the criteria used. But the overall pattern, across many projects and fields, is that published effects tend to be smaller and less reliable than the original literature suggested. The National Academies of Sciences, Engineering, and Medicine (2019) concluded that there was no single rate of non-replicability across science, but that avoidable sources of non-replicability, such as inadequate methods and selective reporting, were common enough to need attention.

Reforms

The crisis prompted a wide set of reforms, many of them grouped under the heading of **open science**. Munafò et al. (2017) set out a manifesto for reproducible science covering methods, reporting, reproducibility, evaluation and incentives. The figure below links the main problems to the reforms designed to address them.

From causes to reforms

What went wrong and what has been done about it

Cause	Effect on the literature	Reform
Low power	Missed effects, inflated estimates	Power analysis, multi-site studies
Flexible analysis	False positives	Preregistration
Questionable practices	Results look stronger than they are	Open data and code, full reporting
Publication bias	Null results disappear	Registered Reports
Incentives	Novelty rewarded over rigour	Valuing replication, new metrics

Drawing on Munafò et al. (2017) and Nosek et al. (2018) ModQu.com

Figure 2. Causes of the replication crisis and the reforms designed to address them. Table by ModQu, drawing on Munafò et al. (2017) and Nosek et al. (2018).

Preregistration. Researchers record their hypotheses, design and analysis plan in a public registry before collecting data. This separates confirmatory tests from exploratory analysis and makes flexible analysis visible (Nosek et al., 2018).

Registered Reports. Journals review a study's question and methods before the results are known and commit to publishing it if the agreed plan is followed (Chambers, 2013). This removes the incentive to produce positive results. Scheel et al. (2021) compared Registered Reports in psychology with standard articles and found that 44 per cent of Registered Reports supported their first hypothesis, compared with 96 per cent of standard articles, a difference consistent with reduced publication bias.

Larger samples and power analysis. Funders and journals increasingly expect studies to justify their sample size, and multi-site collaborations pool data from many laboratories to achieve power that no single team could.

Open data and code. Sharing data and analysis code allows others to check that published results follow from the data, and to test alternative analyses.

Replication as a valued activity. Journals and funders have begun to publish and support replications, which had previously been seen as unoriginal.

Better statistical practice. Researchers are encouraged to report effect sizes and confidence intervals, to treat p-values as one piece of evidence among several, and to distinguish clearly between exploratory and confirmatory findings (Wasserstein & Lazar, 2016).

Changes to incentives. Because the underlying problem lies partly in what careers reward, reformers have argued for evaluating researchers on the quality and openness of their work rather than the number of papers they publish (Munafò et al., 2017). Some funders and institutions now ask applicants to describe their most important contributions, including datasets, software and replications, instead of listing every publication. Changes of this kind are slow, but without them the pressures that Smaldino and McElreath (2016) described will continue to work against better methods.

The reforms are not without costs. Preregistration takes time, larger samples cost money, and open data raises questions of privacy and consent that must be handled carefully. Exploratory research, which generates new ideas, remains valuable and should not be discouraged; the aim is to label it honestly, not to stop it. The direction of change is nonetheless clear: toward research whose results can be checked, repeated and trusted.

Warning signs in a single study

The research on replication suggests several features that should make a reader more cautious about any single finding, inside or outside academia:

- **A small sample**, especially combined with a large reported effect.
- **A p-value just below 0.05**, or several results that all fall just below the threshold.
- **A surprising claim** that conflicts with a large body of previous evidence.
- **Many possible outcomes or subgroups**, with only a few reported.
- **No preregistration or analysis plan**, so that the reader cannot tell which tests were planned.
- **No data or materials available** for others to check.
- **No independent replication**, particularly when the finding comes from a single team.

None of these proves that a finding is wrong. Each raises the probability that it will not hold up, and several together are a strong reason to wait for further evidence before acting.

Lessons for business and applied research

The replication crisis is often discussed as a problem of academic psychology, but its lessons apply wherever data are used to support decisions. Commercial research has its own pressures: clients want clear answers, analysts face deadlines, and a dramatic finding is more welcome than a null one. A marketing test analysed in many ways until one shows a significant uplift, or a survey reported only on the questions that support a preferred conclusion, suffers from the same problems as a flexible academic study.

The practical remedies are the same as well. Agree the key questions, outcomes and analysis before the data arrive. Make sure samples are large enough to detect effects that matter. Report all the results, not only the favourable ones. Treat a striking single finding as a hypothesis to test again, not a conclusion to act on. And keep data and analysis files so that results can be checked.

Conclusion

The replication crisis revealed that many published findings were less reliable than they appeared, because of small samples, flexible analysis, selective publication and incentives that rewarded novelty over rigour. The crisis has not shown that research cannot be trusted. It has shown that the reliability of research depends on how it is done and reported, and it has produced practical reforms, from preregistration to open data, that make results more trustworthy. For anyone who uses evidence, the lasting lesson is to ask not only what a study found, but whether it would be found again.

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452–454. https://doi.org/10.1038/533452a

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*(7391), 531–533. https://doi.org/10.1038/483531a

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407–425. https://doi.org/10.1037/a0021524

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & 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

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at *Cortex*. *Cortex, 49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*(3), 145–153. https://doi.org/10.1037/h0045186

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE, 5*(4), Article e10068. https://doi.org/10.1371/journal.pone.0010068

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

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

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science, 351*(6277), 1037. https://doi.org/10.1126/science.aad7243

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, Article 0021. https://doi.org/10.1038/s41562-016-0021

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. National Academies Press. https://doi.org/10.17226/25303

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

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

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530. https://doi.org/10.1177/1745691612465253

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*(9), 712. https://doi.org/10.1038/nrd3439-c1

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science, 4*(2). https://doi.org/10.1177/25152459211007467

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

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*(9), Article 160384. https://doi.org/10.1098/rsos.160384

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108