

From the research bench

What is publication bias?

Publication bias is when a study's result decides whether it reaches print. In one known set of studies, most null results were never even written up.

Author	Published	Reading time	Words
Sinan Isoglu, MBA (Quantic) Commercial growth leader, lecturer and doctoral researcher	18 October 2026	16 min	3,498

Management summary Publication bias is the selection of studies for publication by the direction or significance of their results; the studies left out sit in what Rosenthal called the file drawer. This concept guide defines it (Head et al.; Munafò et al.; Franco et al.), separates it from p-hacking, reports Sterling's count of four psychology journals in which 286 of 294 articles using significance tests rejected the null hypothesis, and Franco et al.'s follow-up of 221 survey experiments, in which 64.6 per cent of the null results were never written up, explains Rosenthal's tolerance for null results, reported as the fail-safe N, and what it leaves unanswered, and gives a test register for a commercial analytics team. The sources are a count of published articles, a follow-up of one research programme, a method note, a sales meta-analysis that applies the method and two perspective articles.

Contents	§ 1	What is publication bias?	2
	§ 2	What did Sterling count in four psychology journals?	3
	§ 3	How large was the file drawer in a known set of studies?	4
	Figure 1	What became of 221 survey experiments, by the strength of their results	5
	§ 4	Did editors or authors fill the drawer?	6
	§ 5	How many hidden studies would overturn a result?	7
	§ 6	What does a large fail-safe N leave unanswered?	8
	§ 7	What does a file drawer look like inside a company?	9
	Figure 2	A test register	10
	§ 8	How far do these findings travel to company test results?	11
		References	12

What is publication bias?

§ 1 of 8

It is the selection of studies for publication by what they found. Franco et al. (2014)¹ open with a definition they quote from an earlier article (their reference 1): publication bias occurs when the publication of study results depends on “the direction or significance of the findings”. One form of it, they write, is that statistically significant results are more likely to be published than insignificant ones of the same research quality, and selection of this kind makes it more likely that published results are Type I errors, false positives, rather than estimates of the true effect, and pushes published effect sizes upwards.

¹ Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505.

For the name of where the rest go, Franco et al.¹, Head et al. and Munafò et al. all cite Rosenthal (1979)⁴: “The extreme view of the ‘file drawer problem’ is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results” (p. 638). He opens by noting that researchers and statisticians had long suspected that the studies published in the behavioural sciences are a biased sample of the studies actually carried out.

⁴ Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.

The term is used in two ways. Head et al. (2015)² name “two widely recognized types of researcher-driven publication bias”: selection, “the file drawer effect”, in which studies with nonsignificant results are published at lower rates, and inflation, which is p-hacking (p. 1). Munafò et al. (2017)³ use the term for the first type alone: publication bias, “also known as the file drawer problem”, “refers to the fact that many more studies are conducted than published” (p. 3). They add that studies with positive and novel results are more likely to be published than studies with negative results or replications, so that the published literature shows “stronger evidence for findings than exists in reality” (p. 3).

² Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3), e1002106.

3 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(1), 0021.

In this guide, publication bias means the first type: whether a study is published depends on the direction or statistical significance of its result, so that the published studies are a selected sample of those conducted, and the others stay in the file drawer. The selection can be made by an author who does not write a study up or does not submit it, or by the editors and reviewers who reject it, as Franco et al.¹ note; the definition does not depend on who makes it or on intent. P-hacking, the second type on Head et al.'s usage, works inside a study, on which analysis is reported, and is left to its own guide. Publication bias is a selection mechanism at the level of whole studies; the selection bias guide covers how the mechanism that puts units into a sample can change the answer in general.

What did Sterling count in four psychology journals?

§ 2 of 8

An almost unbroken run of significant results. Sterling (1959)⁵ took every issue of one year (1955, or 1956 for one of them) of four psychology research journals, chosen at random from four major areas of psychology. Of 362 research reports, 294 used tests of significance. Of those 294, 286 rejected the null hypothesis and 8 did not: 97.28 and 2.72 per cent. The criterion was the 5 per cent level, except for two reports at the 10 per cent level that Sterling counted as rejections because their authors had expressly said so (p. 32). None of the 362 reports was a replication of a previously published experiment. As a check, he drew 100 research articles at random from Psychological Abstracts for 1952 to 1957: 94 reported positive results, 5 negative results and one was a replication (pp. 31 to 32).

5 Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance, or vice versa. *Journal of the American Statistical Association*, 54(285), 30-34.

He did not assume that explicit editorial rules produced the pattern. The single largest factor, he suggested, might be an implicit agreement among authors, and he used the label publication policy largely for convenience; what the tables show is what gets printed. His argument is about what follows. Where research with nonsignificant results is not published, other investigators do not know it was done, so it may be repeated independently until a significant result occurs by chance, a Type I error, and that result is published; significant results in such

fields are seldom verified by independent replication; so the literature of a field may consist in substantial part of false conclusions (p. 30). For the reader, therefore, “The risk stated by the author cannot be accepted at its face value” once the conclusions are in print: the stated 5 per cent applies to one experiment, while the reader may be looking at the one experiment among similar ones that obtained a large difference by chance (p. 34).

Sterling’s⁵ premises were that experiments rejecting the null hypothesis are printed with a greater probability than those that do not, that a printed experiment is very unlikely to be replicated, and that many more experiments are performed than appear in journals. He called them substantive and wrote that strong supporting evidence for them, “beyond that given here, is hard to come by” (p. 33). He added that a few minutes of browsing journals in biology, chemistry, medicine, physiology and sociology “show that the same usages are widespread” in other sciences (p. 34); that was an impression, not a count. A count of journals shows what was printed; it cannot show what was run and left out. The [reproducibility guide](#) separates rebuilding a result from its own inputs from replicating it with new data, the step Sterling found missing.

How large was the file drawer in a known set of studies? § 3 of 8

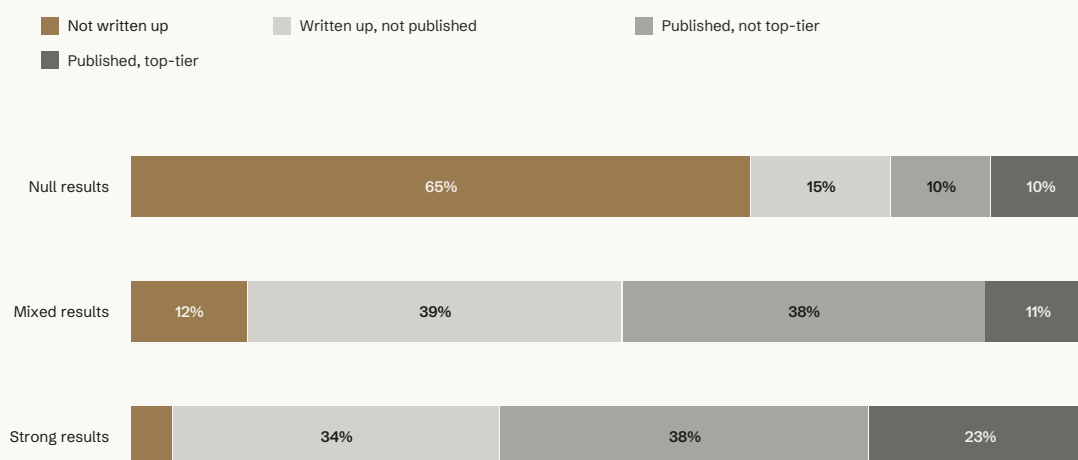
More than half of the studies had not appeared in a journal, and most of the null results were never even written up. Franco et al. (2014)¹ found what Sterling lacked: a set of studies whose unpublished members could be counted. They used TESS (Time-sharing Experiments in the Social Sciences), a programme sponsored by the US National Science Foundation and established in 2002, in which researchers propose survey experiments to be run on nationally representative samples, and TESS peer reviews the proposals and awards the grants competitively. Because TESS requires adequate statistical power, a study that failed to reach significance did not fail simply for lack of sample size (p. 1 of the Science Express version).

They followed 221 studies run between 2002 and 2012, 89 per cent of the 249 in the archive; seven published as book chapters and 21 with unknown status or results were excluded. They searched for every publication and emailed the authors of over 100 studies they could not trace. Results were classified by how the authors themselves described them: strong when all or most hypotheses were supported, null when all or most were not, mixed otherwise. 41 per cent of the studies reported strong results, 37 per cent mixed and 22 per cent null;

80 per cent had been written up and 48 per cent published in academic journals (pp. 1 to 2 of the Science Express version). Figure 1 shows what became of each group.

Figure 1 What became of 221 survey experiments, by the strength of their results

Of the studies with null results, 64.6 per cent were never written up; of those with strong results, 4.4 per cent.



Each bar is one result group and sums to 100 per cent; the bars hold 48, 82 and 91 studies, this guide's sums of the analysed counts in Table 2 (p. 6 of the Science Express version). Segment labels are each value's share of its bar, rounded to whole per cent; the table prints one decimal, its strong column sums to 100.1, so 38.5 is labelled 38, and the 4.4 per cent of strong results never written up is too narrow to label. Results are classed by the authors' own description.

Source Franco et al. (2014), Table 3, p. 6 of the Science Express version held (pp. 1 to 6, published online 28 August 2014): column percentages of the 221 TESS survey experiments conducted from 2002 to 2012, by the strength of results their authors reported and by publication status. Book chapters and studies with an unknown status or result are excluded, as in the authors' analysis. The authors' figures, not a re-estimation.

Strong results were 40 percentage points more likely to be published than null results and 60 percentage points more likely to be written up. About 20 per cent of the studies with null results appeared in print, against roughly 60 per cent of those with strong and 50 per cent of those with mixed results; studies with null findings were more than 20 per cent of the sample but less than 10 per cent of the published articles based on it. A test of independence between result and

publication status is easily rejected ($\chi^2(6) = 80.3, p < 0.001$; pp. 1, 2 and 6 of the Science Express version).

Did editors or authors fill the drawer?

§ 4 of 8

Mostly the authors, in this set: the selection by result happened at the write-up, before review. “Few null findings ever make it to the review process”, Franco et al.¹ write. 64.6 per cent of the studies with null results were never written up, and among the studies that were written up, the authors found no significant relationship between the strength of results and publication status (p. 2 of the Science Express version). Their summary is that “authors do not write up and submit null findings” (p. 1 of the Science Express version). It agrees with the earlier work they cite, in which analyses of clinical registries and of abstracts submitted to medical conferences find “little to no editorial bias” against studies with null findings, and failure to publish appears to be most strongly related to authors’ perceptions that null results are uninteresting (p. 1 of the Science Express version).

They asked why. Of 26 detailed email replies from researchers whose studies yielded null results and who did not write a paper, 15 said they abandoned the project because they believed that null results have no publication potential, even when they found the results interesting personally, and 9 gave writing up a lower priority and turned to other projects. Two authors whose studies “didn’t work out” later published papers supporting their initial hypotheses with findings from smaller convenience samples (p. 2 of the Science Express version). From their communications with the authors of many experiments with null findings, Franco et al.¹ conclude that some researchers anticipate “the rejection of such papers” while many simply lose interest in projects they count as unsuccessful (p. 2 of the Science Express version). The decision not to write up was the authors’, then, but for some of them it was a forecast of what editors and reviewers would do.

The authors say that TESS studies are “clearly not a random sample” of social science research, and argue that publication bias elsewhere is unlikely to be less severe, because TESS data were costly to collect and their authors should, if anything, be more motivated to publish null results. The strength of each result rests on its authors’ own report, and the authors’ sensitivity analysis finds their results robust to even dramatic and unrealistic rates of misclassification. Of the remedies already proposed, two-stage review, pre-analysis plans and pre-registration, they write that these should be complemented by incentives not to “bury insignificant results in file drawers” (p. 2 of the Science Express version). Munafò et al.³ (2017) make a related point for

the strongest form of pre-registration, which registers the study with a commitment to make the results public and fixes its design and analysis in advance: “In principle, this addresses publication bias by making all research discoverable, whether or not it is ultimately published” (p. 3).

How many hidden studies would overturn a result?

§ 5 of 8

Rosenthal (1979)⁴ asked the question a reviewer can answer without finding the drawer: “For any given research area, one cannot tell how many studies have been conducted but never reported” (p. 638). Instead of counting them, he computed how many studies averaging a null result would have to be in the file drawers before the combined result of the studies the reviewer has felt to just significant, at $p = .05$. He called this number the tolerance for future null results; if a few more null results would bring the combined result down that far, the finding “is not resistant to” the file drawer threat (p. 639). Verbeke et al.⁶ (2011), in a meta-analysis of what drives salesperson performance, calculated it as the fail-safe N , to assess the plausibility of a file drawer problem: the number of unlocated studies “with null results needed to reduce the cumulated effect across studies to the point of non-significance” (p. 413).

⁶ Verbeke, W., Dietz, B., & Verwaal, E. (2011). Drivers of sales performance: A contemporary meta-analysis. Have salespeople become knowledge brokers? *Journal of the Academy of Marketing Science*, 39(3), 407–428.

The calculation uses the method of adding Z s. Each study’s p -value is turned into a standard normal deviate Z , and the Z s of the k studies are added and divided by the square root of k , which gives a combined Z . Setting that combined Z , with X extra studies at $Z = 0$ added, equal to 1.645 and solving gives Rosenthal’s⁴ alternative formula, $X = (\text{sum } Z)^2 / 2.706 - k$, where 2.706 is 1.645 squared; his Equation 3 states the same with the mean Z (p. 639). In his first example, 94 experiments on interpersonal self-fulfilling prophecies had a mean Z of 1.014, and 3,263 studies averaging null results would have had to sit in file drawers to overturn them; a later summary of 311 studies, with a mean Z of 1.180, gave 49,457. In the other direction, 15 studies averaging $Z = 0.50$ have a combined p of .026, and 6 null studies in a drawer would make it nonsignificant; 2 studies averaging $Z = 2.00$ give p of about .002, and 4 would (pp. 639 to 640).

How many is too many to be plausible, he left open: no firm guidelines could yet be given, he wrote, for what number of unretrieved or

unpublished studies is unlikely. In some areas 100 or even 500 unpublished studies might be plausible, in others even 10 or 20 seems unlikely. He suggested, tentatively, treating a result as resistant when the tolerance reaches $5k + 10$, a level he thought conservative but reasonable, and that reviewers report the tolerance beside every combined p (p. 640).

What does a large fail-safe N leave unanswered?

§ 6 of 8

Three things, on this guide's reading of the sources. First, it concerns significance, not size. Rosenthal's⁴ question is whether the effect is "real," that is, "not a Type I error" (p. 639), and Verbeke et al.⁶ define the number by the point of non-significance. It says nothing about how far the published effect sizes overstate the true ones, the consequence Franco et al.¹ name when they describe selection by significance as "biasing effect sizes upwards" (p. 1 of the Science Express version): a literature can need thousands of hidden studies to lose its significance and still report effects that are too large.

Second, it rests on assumptions Rosenthal states. The studies in the drawer are taken to average a null result, $Z = 0$, and his equations assume that each of the k studies is independent of the others (p. 639). If the unpublished studies leaned against the effect instead, fewer of them would overturn it; that follows from the arithmetic and was not tested in the article. Third, the threshold is a judgement: Rosenthal gives no firm guideline and offers $5k + 10$ as one reasonable level. The [source verification guide](#) traces a citation to the identity, version and held text of the work; publication bias concerns which works reached print at all, so a source can pass every step there and still be one of the studies that were selected for print.

What does a file drawer look like inside a company?

§ 7 of 8

It can look like a quarterly review that lists only the tests that won. The sources concern journals, so this section is the author's reasoning. A pricing or sales analytics team can run many tests in a year: price points by region, discount bands, renewal terms, a new sales playbook in some territories. Someone decides which of them reach the review. If tests with null results are not written up and not shown, the review becomes a dashboard of winners, and the team's own record carries the selection Franco et al.¹ found among academic authors, at the same step: the write-up, which the analyst controls and which, as in TESS, can anticipate what the review is expected to want. Sterling's⁵ chain can run inside a team as well: a null test that leaves no note may be run again later, and a rerun that crosses the threshold by chance is the one that gets reported.

A company has one advantage over a journal reader: it can count its own file drawer. TESS gave Franco et al. a known population of studies; a register of every test, made before the test starts, gives a team the same. The review can then show how many tests were run, how many came out strong, mixed or null, and how many of each reached the report, and Rosenthal's⁴ question, how many hidden studies it would take to overturn a result, need not be asked, because the number of tests left out is known. The uplift specification essay shows, for nudge trials, how far published effects can sit from a census of trials run by nudge units, and why a claim should name the evidence it rests on.

Figure 2 A test register

Register every test before it starts, then record its result, its write-up and whether it reached the report.

Test	Planned size	Result	Write-up	Reported	If left out
Question, and date registered?	Sample or run, and effect sought, fixed before start?	Strong, mixed or null, by plan?	Note written, and filed where?	Which review shows it?	Why was it left out?

Next to any list of winners, show how many tests were run and how many came out null.

One row per test, written before the test starts. The planned size, read with the effect sought, shows whether a test that came out null could have detected the effect the team cared about; the last three columns show where the file drawer is.

Source Author's decision worksheet. The register as a known population follows Franco et al. (2014), pp. 1 and 2 of the Science Express version; the planned size field follows their note that TESS requires adequate statistical power (p. 1), and the effect sought beside the size is this guide's addition; the write-up and report fields follow their finding that most null results were never written up (pp. 1 and 2, Table 3 on p. 6) and Munafò et al. (2017), p. 3; the reason field follows the authors' explanations Franco et al. report (p. 2). The fields are reader inputs; no company data is supplied.

An illustration, with no company data: a pricing team registers 24 price tests in a year, each with its question, its sample, its run length and the smallest effect worth finding fixed before the start. Seven show a significant increase in revenue per account. Without the register, the quarterly review would list the seven. With it, the review shows all 24 by result, each null test has a short note filed where the next analyst will look before proposing the same test, and any planning figure built on the seven is read beside the count of tests that came out null. The planned size matters as well: a null result from a test too small to detect the effect the team cared about says little either way; Franco et al.¹ could rule that out for TESS because the programme required

adequate statistical power. These are the author's examples and have not been tested.

How far do these findings travel to company test results?

§ 8 of 8

Not far as measured numbers. Sterling (1959)⁵ counted four psychology journals in 1955 and 1956: his count shows what was printed, and his account of why rests on assumptions he called substantive. Franco et al.¹ (2014) followed one programme of peer-reviewed, adequately powered survey experiments in the social sciences, with results classed from the authors' own reports. Rosenthal (1979)⁴ gives a calculation under stated assumptions, not a measurement of any drawer, and Verbeke et al. (2011)⁶ apply it in a meta-analysis whose subject is salespeople, not publication bias. Head et al. (2015)² and Munafò et al. (2017)³ are perspective articles, used here for how they use the term and, in Munafò et al.'s case, for what they expect of pre-registration. None of them observed a company's test programme.

On this guide's reading, two points carry over with some confidence, because neither depends on the sample: a set of results selected for being significant overstates the evidence, and a file drawer can fill before any reviewer sees the work, at the decision about what gets written up, as most of it did in the one set of studies here whose unpublished members could be counted. Whether an analytics team's file drawer looks like the one in TESS, and whether a test register changes what a team reports or decides, the sources do not say.

 End of essay

References

- 1 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>
- 2 Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3), e1002106. <https://doi.org/10.1371/journal.pbio.1002106>
- 3 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(1), 0021. <https://doi.org/10.1038/s41562-016-0021>
- 4 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>
- 5 Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance, or vice versa. *Journal of the American Statistical Association*, 54(285), 30–34. <https://doi.org/10.2307/2282137>
- 6 Verbeke, W., Dietz, B., & Verwaal, E. (2011). Drivers of sales performance: A contemporary meta-analysis. Have salespeople become knowledge brokers? *Journal of the Academy of Marketing Science*, 39(3), 407–428. <https://doi.org/10.1007/s11747-010-0211-8>

How to cite this essay

Isoglu, S. (2026, October 18). What is publication bias? Sinan Isoglu.
<https://isoglu.com/insights/journal/what-is-publication-bias/>
Online version: <https://isoglu.com/insights/journal/what-is-publication-bias/>

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher

© 2026 Sinan Isoglu · isoglu.com · All rights reserved