A selection model shrinks the overstatement in published effects and cannot pick which ones replicate

A preregistered leave-one-field-out test on 1,222 significant originals from the Reproducibility Project Psychology and the FORRT Replication Database

Zain Dana Harper

Version 1.0, 9 November 2026. Finding F172 in the findings ledger. Signal (tier B).

Abstract

Published effects that cleared a significance threshold tend to be overstated, and replications of them succeed less often than their own statistics suggest. We asked whether one model of that filter, fitted only to the significant originals of other fields, can predict which effects replicate in a field it has not seen. On 1,222 significant originals in 307 replication projects, pooled from the Reproducibility Project Psychology and the FORRT Replication Database, each of eight fields was held out in turn and forecasts were scored by Brier score. The selection model beat taking the originals at face value by 0.142 in Brier score (95% interval 0.108 to 0.186) and did not beat the base rate, at -0.004 (-0.039 to 0.027). A small logistic model on the original z value and the two sample sizes beat the selection model by 0.037 (0.011 to 0.069). The preregistered kill rule fired, and the claim that the model tells which effects replicate is withdrawn. The reading that significant originals are inflated survives on these data.

1 What this means for a reader

If a published effect cleared p below 0.05, expect a replication to find it smaller, and expect the chance of a successful replication to be well below what the original's own numbers imply. A model of the significance filter turns that expectation into a number that is far better than trusting the original. It does not tell you which single finding will hold. On these data the plain original z value, read through a small regression, is the stronger tool for picking claims to check.

2 Introduction

The Open Science Collaboration [1] repeated 100 studies from three psychology journals and found that far fewer replications than originals reached significance. Since then, replication results have been gathered into databases. The FORRT Replication Database, FReD, joins earlier collections and pairs original findings with their replications at the level of single effects [4].

One explanation of the gap needs no misconduct. When results are published only if significant, the significance filter lets through estimates that happened to land high. Gelman and Carlin [5] named the two resulting errors: the sign of a significant estimate can be wrong (Type S), and its size can be exaggerated (Type M), and both grow as power falls. A replication of the same size then succeeds less often than the original's estimate predicts.

Two families of tools forecast replication. Statistical models of selection, such as z-curve [6], fit the distribution of significant z values and estimate how often exact replications should succeed. Learned models predict from features of the original; Altmejd and colleagues [7] trained such a model on replications in economics and psychology and found, among other things, that interaction effects replicate worse than main effects.

This paper puts a selection model in a held-out test against three baselines, with all parameters fitted away

from the field being forecast. The question is narrow: does a model of the filter pick out which effects replicate, or does it only set the level? Our search did not find a test of this exact design. That is what the search returned, and it does not show that none exists.

3 Preregistration and design

The design was written as a preregistration and hash-locked on 31 October 2026 at 08:21 UTC, before any replication outcome was scored. Seen before the lock: column names, counts by category, the number of original statistics that could be converted, and outcomes simulated by the code. The published replication rate of the Reproducibility Project, 36 percent, was known, so the forecast for that field was not blind to its headline.

The locked file names the data, the conversions, the model, the held-out scheme, the baselines, the score, the interval and test methods, the planted check, two data checks, each control, the forecasts and the kill conditions. Nothing was added after the run.

Kill conditions, without discretion:

- If the selection model fails to beat the base rate on held-out fields, it predicts a rate and not which effects replicate, and the study-level claim is withdrawn.
- If it fails to beat face value, the winner's-curse reading is dropped for these data.
- If the planted check or a data check fails, the affected verdicts are void.

4 Data and provenance

Two public sets were read through the Open Science Framework's interface: the Reproducibility Project Psychology data [1] [2], released under CC0, and FReD [3], released under CC BY 4.0. Rows of the Reproducibility Project were removed from FReD so that no effect enters twice. Only numeric columns were read from the Reproducibility Project file; names and emails in it were never loaded. Outputs stay at field and project level, and no paper or author is scored or named.

Effects were converted to a correlation from r, d, g, h, standardised beta, odds ratios, eta squared, r squared, f squared, and parsed t, F with one numerator degree of freedom, Z and chi squared with one degree of freedom. Other effect types were dropped. Of 2,164 FReD rows, 1,344 converted. Only originals significant by our conversion entered, since the model conditions on that filter: 201 FReD rows and 18 Reproducibility Project rows were dropped as not significant. That left 1,222 effects, 79 from the Reproducibility Project and 1,143 from FReD, in 307 project clusters, the largest with 156 rows.

Fields are coarse buckets of our own: seven groups of FReD rows (cognitive, economics, management, marketing, other psychology, social and other) and the Reproducibility Project as an eighth. Which effects got replicated was itself selected by the people who chose to replicate them. The results describe this set of replicated effects, not published work at large.

5 Methods

Effects on one scale

Each effect is put on the Fisher z scale, with a standard error from the sample size n:

$$y = \frac{1}{2}\ln\frac{1+r}{1-r} \tag{1}$$

$$s = \frac{1}{\sqrt{n-3}} \tag{2}$$

An original enters only if y/s exceeds 1.96 in absolute value. A replication succeeds if its own y/s exceeds 1.96

in the original's direction. The same rule is applied to every row, whatever the database's own label says.

The selection model

The true effect θ of an original is either exactly zero, with weight π_0, or drawn from a normal distribution with mean zero and spread τ:

$$g(\theta) = \pi_0 \delta(\theta) + (1 - \pi_0)\varphi(\theta; 0, \tau^2) \tag{3}$$

where δ is a point mass at zero and $\varphi(x; m, v)$ the normal density with mean m and variance v. The estimate is normal around the true effect, so an original's density before selection is $f(y) = \int g(\theta)\varphi(y; \theta, s^2)d\theta$. Because only significant originals are seen, each original's likelihood is divided by the chance $w(s)$ that an estimate with its standard error would be significant:

$$L = \prod_i \frac{f(y_i)}{w(s_i)} \tag{4}$$

The parameters π_0 and τ are fitted by maximum likelihood on the training fields. The weight π_0 is a mixture weight under an assumed prior. It is not the share of false hypotheses, and it measures no misconduct or intent.

Forecasts

For original i, let $h_i(\theta)$ be the density of the true effect given the original's estimate, and $\psi_i(\theta)$ a replication's statistical power at its own sample size, meaning the chance that it succeeds by the rule above when the true effect is θ. The forecast is

$$q_i = \int \psi_i(\theta) h_i(\theta) d\theta \tag{5}$$

Five forecasts compete with it, each built only from training rows. The face-value forecast sets θ equal to the original estimate. The base-rate forecast gives every row the training success rate. A logistic regression uses the original z value and the logs of the two sample sizes. A normal prior with no spike drops the point mass. A transport term lets the replication's true effect differ from the original's by a normal amount whose variance h is fitted on training replications.

Score and held-out scheme

Forecasts are scored by the Brier score over N held-out rows with outcomes o_i of 1 or 0:

$$B = \frac{1}{N} \sum_i (q_i - o_i)^2 \tag{6}$$

In a leave-one-field-out test each of the eight fields is forecast with every parameter fitted on the other seven, and the held-out scores are pooled. Skill over the base rate is $B_{base} - B_{sel}$, so a positive value favours the selection model.

Intervals and tests

Each comparison is a difference in pooled Brier score, the competitor's minus the selection model's. Its 95% interval comes from a project-cluster bootstrap with 2,000 draws. Its p value comes from a sign-flip permutation test over projects with 4,000 draws, and the four comparisons are adjusted together with the Benjamini-Hochberg correction. The prior's shape was compared by BIC on all originals against a single normal and a two-normal mixture.

6 Controls and planted checks

The planted check used the real sample-size pairs with outcomes simulated from the model at three known settings of π_0 and τ: 0.5 and 0.25, 0.05 and 0.35, 0.8 and 0.3, five datasets each. Across the 15 planted datasets the fit recovered π_0 with mean absolute error 0.051 and the success rate with mean error 0.009, and the selection model beat face value in every one. The rules were an error below 0.1, a rate error below 0.03 and a win in at least 90 percent. The fit was weakest at π_0 of 0.05, where two of five datasets gave estimates near 0.18.

Two data checks tested the conversions. The Reproducibility Project rows we kept carried 34 reported successes against the 35 published, within the rule of 4. FReD rows labelled successful by the database had a success rate by our rule of 0.86, and rows labelled failed 0.15, a gap of 0.71 against a rule of at least 0.3. By our own uniform rule, 27 of the 79 kept Reproducibility Project rows succeeded.

Two controls give the null distribution of skill over the base rate. Shuffling replication outcomes across rows (500 draws) gave a 95th percentile of -0.110. Shuffling original statistics across rows while keeping each replication's size (200 draws) gave -0.092 and measures how much skill comes from replication size alone. Both sit far below zero, because a confident forecast paired with the wrong outcome is punished hard. Beating them shows that the real pairing carries signal. The base-rate comparison is the test that decides.

7 Results

Table 1 and Figure 1 give the pooled held-out Brier scores. The selection model at 0.260 sits next to the base rate at 0.256, far ahead of face value at 0.402 and behind the logistic model at 0.223.

Table 1. Held-out Brier scores on 1,222 effects, pooled over the eight held-out fields. Lower is better. The project-weighted column gives each project equal weight.

Forecast	Brier	Project-weighted
Logistic regression	0.223	0.231
Selection model with transport term	0.232	0.234
Base rate	0.256	0.256
Selection model	0.260	0.244
Normal prior, no spike	0.345	0.344
Face value	0.402	0.426

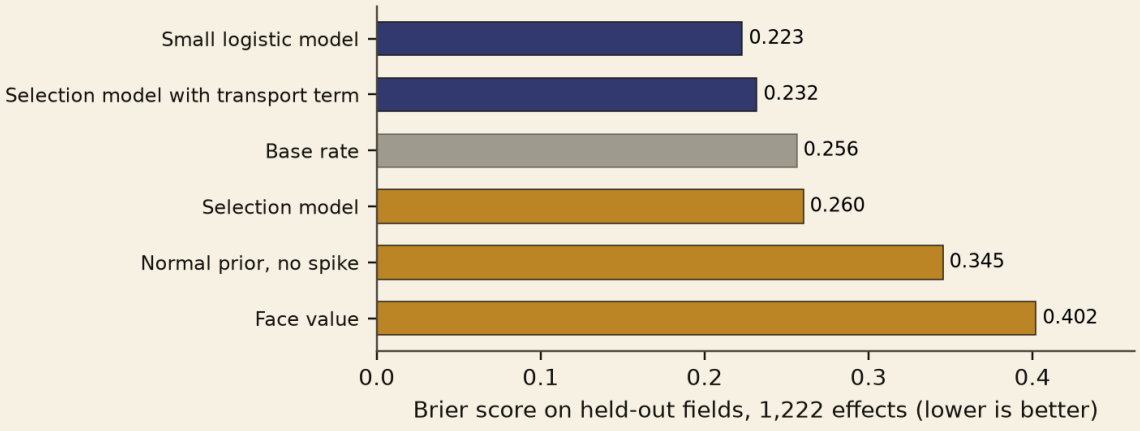


Figure 1. Held-out Brier scores of the six forecasts. Navy bars beat the base rate (grey); ochre bars do not. Drawn from the run's result file.

Table 2 gives the four preregistered comparisons. The selection model beat taking the originals at face value by 0.142 in Brier score (95% interval 0.108 to 0.186) and did not beat the base rate, at -0.004 (-0.039 to 0.027). A small logistic model on the original z value and the two sample sizes beat the selection model by 0.037 (0.011 to 0.069).

Table 2. Competitor's Brier score minus the selection model's, so a positive value favours the selection model. Interval from the project-cluster bootstrap, 2,000 draws; p from 4,000 project sign flips; q after the Benjamini-Hochberg correction over the four.

Comparison	Difference	95% interval	p	q
Face value	0.142	0.108 to 0.186	0.00025	0.0005
Base rate	-0.004	-0.039 to 0.027	0.811	0.811
Logistic regression	-0.037	-0.069 to -0.011	0.010	0.014
Normal prior, no spike	0.085	0.059 to 0.118	0.00025	0.0005

By field (Table 3, Figure 2), skill over the base rate was positive in three of the six fields the run counted as large, short of the preregistered target of 75 percent. The model's predicted rate sat above the observed rate in seven of the eight fields and below face value in all eight. For the Reproducibility Project the model predicted a success rate of 0.50 against 0.34 observed, where face value predicted 0.83; in social psychology it predicted 0.65 against 0.41, where face value predicted 0.89.

Table 3. Held-out results by field. Rates are shares of rows that replicated by our rule. Skill is the base rate's Brier score minus the selection model's.

Field	Rows	Projects	Observed	Predicted	Face value	Skill
Cognitive	200	50	0.50	0.63	0.88	0.020
Economics	16	7	0.50	0.50	0.79	-0.006
Management	82	4	0.79	0.84	0.96	0.088
Marketing	84	6	0.37	0.66	0.98	-0.053
Other	163	43	0.62	0.69	0.89	-0.002
Other psychology	229	70	0.62	0.72	0.98	0.016
Reproducibility Project	79	79	0.34	0.50	0.83	0.048
Social	369	83	0.41	0.65	0.89	-0.050

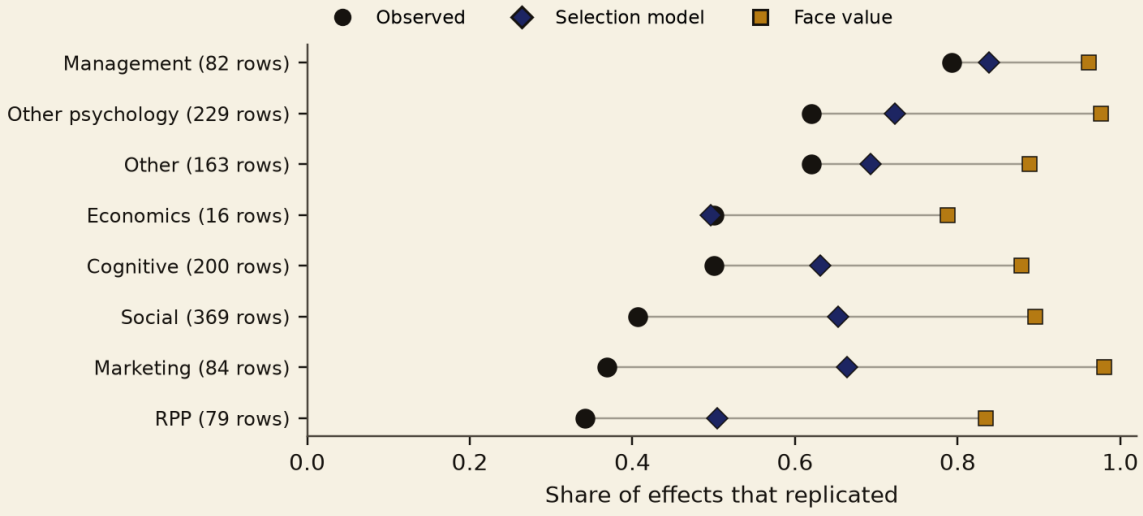


Figure 2. Observed replication rate (circles) against the selection model's forecast (diamonds) and face value (squares), by held-out field. The model moves every field's forecast toward the observed rate and stops short of it in most.

The forecasts were also miscalibrated at the top, as Figure 3 shows. In five bands of equal count, calibration was close in the lowest two and overconfident above: effects given a mean chance of 0.97 replicated at 0.57, and effects given 1.00 replicated at 0.83.

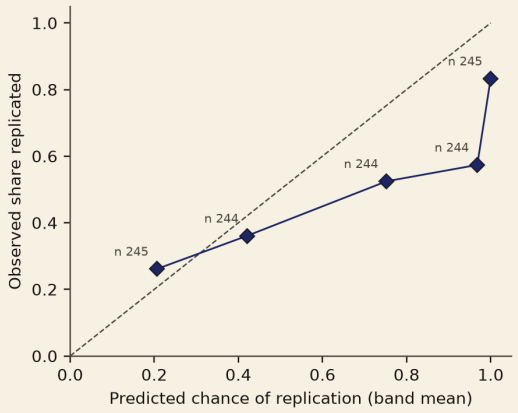


Figure 3. Calibration of the selection model's held-out forecasts in five bands of 244 or 245 effects. The dashed line is perfect calibration.

The prior's shape did not hold up. The spike-and-normal prior beat a single normal by 192 in BIC (4,946.8 against 5,138.4), passing its rule of 10. A two-normal mixture beat the spike by 289 (4,657.5). On all originals the fitted π_0 was 0.82 and τ 0.51, and that description is a poor one. The transport variance h hit the top of its grid, 1.0 on the Fisher z scale, in every fold.

On the 701 rows whose effect type carries a sign, the selection model scored 0.273, face value 0.402 and the base rate 0.256. The ranking of these three is the same as in the full set. With each project weighted equally, a preregistered sensitivity analysis with no decision rule attached, the selection model scored 0.244 against the base rate's 0.256 (Table 1), so the point comparison with the base rate changes sign when projects are weighted equally. No interval was computed for that version, and the preregistered verdict uses the row-weighted score.

Real skill over the base rate was -0.004, above the 95th percentiles of both shuffle controls (-0.110 and -0.092; p 0.002 and 0.005). Table 4 sets the forecasts written before the run against the outcomes.

Table 4. Forecasts fixed in the preregistration against what the run gave.

Quantity	Forecast	Outcome
Brier, selection model	0.21	0.260
Brier, face value	0.27	0.402
Brier, base rate	0.235	0.256
Brier, logistic	0.215	0.223
Fitted pi0 and tau	0.5 and 0.3	0.82 and 0.51
Reproducibility Project predicted rate	0.40	0.50 (observed 0.34)
Beats face value	probability 0.9	yes
Beats base rate	probability 0.7	no
Beats logistic	probability 0.3	no
Spike beats normal by BIC 10	probability 0.5	yes
Share of fields with positive skill	0.6	0.5
Share at least 0.75	probability 0.35	no
Skill above shuffled-original control	probability 0.6	yes

8 Discussion

The result splits the selection reading in two. The part about level holds: forecasts that ignore the filter are

badly overconfident, and modelling the filter brings the Brier score down from 0.402 to 0.260. This is the pattern Gelman and Carlin [5] describe, measured here as forecast error on held-out fields.

The part about picking effects fails. A forecast that ignores every effect's statistics and uses the training success rate does as well. On Brier score, the model's spread of forecasts across effects adds nothing over the base rate, and its top bands are overconfident. The logistic model, which uses the same original statistics without a model of selection, does better. A selection model with a rigid prior shape spends its freedom on the level and leaves the ranking to the prior, and the prior fits poorly by BIC.

The transport term points the same way. Allowing the replication's true effect to drift from the original's improved the score to 0.232, and its variance ran to the top of the grid in every fold. The gap between originals and replications on these data is larger than the selection model with a fixed true effect accounts for. Hidden moderators, differences in fidelity, real change over time and design differences ignored by the Fisher z standard error could all add to it, and the original statistics cannot tell them apart.

For practice, this narrows what a selection model is good for. It can set a sober expected replication rate for a body of work. Altmejd and colleagues [7] and z-curve [6] aim at related targets by other routes; this paper does not compare against them.

9 Limitations

- An earlier summary of ours said the model corrects the replication rate. Read against the table, it only shrinks the gap: the Reproducibility Project forecast missed by 0.16 against a preregistered tolerance of 0.05.
- The two-normal prior fitted better by BIC, so π_0 of 0.82 should not be read as a share of false hypotheses. The two-normal prior was compared by BIC only and was not tested on held-out fields.
- Fisher z standard errors ignore study design, so within-person studies look noisier than they are. Unsigned effect types assume the same direction in original and replication. Effects were converted from mixed metrics, and beta was treated as r.
- FReD has large nested projects; one has 156 rows. Management has 82 rows from 4 projects.
- Fields are our own coarse buckets.
- The Reproducibility Project forecast was not blind to its published rate, and the data's category counts were seen before the lock.
- The model predicts from original statistics. It cannot say why an effect failed, and a forecast on one effect is no verdict on that effect, its authors or their field.
- Scoring was done by our own code and was not blind.

10 Verified

Verified: on 9 November 2026 the row, project and conversion counts, the six Brier scores and their project-weighted versions, the four comparison differences with intervals, p and q values, the per-field rows, projects, rates and skills, the five calibration bands, the planted and data-check values, the BIC values, the fitted prior and transport values, the signed-row sensitivity, both shuffle quantiles and all nine verdicts were re-read from the run's result file. They match the findings ledger. The forecasts were read from the preregistration. The preregistration's SHA-256 was recomputed on that date and matches the hash in the result file. All three figures were drawn by a script that reads only that result file. The study was not re-run for this paper. The literature in the introduction was found by searches run on 9 November 2026 and read at the level of abstracts, publisher pages and search summaries; the full texts were not read. A Verified line lists what was checked, and anything not listed was not checked.

11 Corrections

- 31 October 2026: the original heading and kill line in our findings ledger said the model corrects the rate. Read against the table it only shrinks the overstatement. This paper states the narrower reading throughout.
- 9 November 2026: first published version.

12 Reproduction

- Data: download the Reproducibility Project Psychology data and FReD from the Open Science Framework, and remove the Reproducibility Project rows from FReD.
- Convert effects to Fisher z with eqs. 1 and 2, and keep originals with y/s above 1.96 in absolute value.
- Fit the prior of eq. 3 by the selection likelihood of eq. 4 on seven fields, forecast the eighth with eq. 5, and repeat for every field.
- Score with eq. 6; compare forecasts with a 2,000-draw project bootstrap and 4,000 project sign flips, adjusted by Benjamini-Hochberg.
- Preregistration SHA-256: 93132f995daad75ca338b5ac8442d262da99b52b5c59d7f1dbfa83fcba3a466f. Plan SHA-256: 1a6cc22eb03f2b7a168e32ecd696a312affde7a8b1da74809a97434e3f9c7b33. Run 20261031T082124Z.
- Cost: the analysis took 13 seconds and both controls under one second on one desktop computer, as timed in the result file.

13 Appendix A. The preregistered verdicts

Each verdict is a set of rules written before the run. A verdict passes only when every rule holds. Values are read from the run's result file.

Table 5. Preregistered verdicts and the rules that decided each one.

Verdict	Deciding rule	Value	Holds
Planted check	pi0 error below 0.1; rate error below 0.03; beats face value in at least 0.9 of datasets	0.051; 0.009; 1.00	yes
Data check	reported Reproducibility Project successes within 4 of 35; FReD label gap at least 0.3	1; 0.71	yes
Beats face value	lower interval bound of the difference above 0	0.108	yes
Beats base rate	lower bound above 0; skill above shuffled-outcome 95th percentile (-0.110)	-0.039; -0.004	no
Originals add skill	skill above shuffled-original 95th percentile (-0.092)	-0.004	yes
Beats logistic	lower interval bound of the difference above 0	-0.069	no
Spike needed	BIC of spike minus normal below -10	-191.6	yes
Most fields show skill	share of large fields with positive skill at least 0.75	0.5	no
Reproducibility Project rate	forecast error at most 0.05	0.163	no

14 Glossary

Each term is defined where it is first used in the text. The same definitions, with the field each comes from and what each excludes, are collected here.

- **significance filter (winner's curse)**. The selection of published results by a significance threshold, which lets lucky overestimates through more often than underestimates, so the published effects of a set of studies are on average larger than the true ones. Field: metascience and statistics. Excludes deliberate misconduct, which the filter neither requires nor detects.
- **selection model**. A model of observed results that writes down both a distribution of true effects and the

rule by which results get seen, here the rule that only originals with z above 1.96 in absolute value are seen, and fits both together. Field: meta-analysis and statistics. Excludes a model that takes the observed results as an unselected sample.

- **preregistration**. A dated written plan for a study, fixed and hash-locked before any result is seen: the question, the data, the statistic, the controls, the forecasts and the kill conditions. Field: research method (meta-science). Excludes a plan revised after looking at the outcome, and a protocol written only for the record after the run.
- **planted check**. Data built so that the right answer is known, run through the same analysis before the real data, to show that the analysis can see the effect when it is there. Field: research method; called a positive control in experimental science. Excludes any check made on the real data, which cannot show what the analysis misses.
- **control**. A version of the data with the structure under test removed (for example shuffled or pooled), scored the same way as the real data, so the real score has something to beat. Field: statistics and experimental design. Excludes a comparison with a different dataset or with a published number.
- **kill condition**. A rule written in the preregistration that, when met, decides against the claim with no room for judgement after the run. Field: research method (our usage, after Popper's falsification). Excludes reasons found after the run for setting a result aside.
- **verdict**. The pass or fail of one preregistered rule set against the run's numbers, read off by code. Field: research method (our usage). Excludes an interpretation written after reading the numbers.
- **void**. The verdict of a run whose planted check failed or whose sample fell below its stated minimum; it says nothing for or against the claim. Field: research method (our usage). Excludes a null result, which comes from a working instrument.
- **Fisher z**. The transform of a correlation r to half the natural log of (1 + r) over (1 - r), on which the sampling error is close to normal with standard error 1 over the square root of n minus 3. Field: statistics. Excludes the z statistic of a test, which is an effect divided by its standard error.
- **face-value forecast**. A forecast that takes each published original effect as the true effect and predicts replication from the power a replication of the given size would have against it. Field: forecasting, our usage. Excludes any correction for how the original was selected.
- **base-rate forecast**. A forecast that gives every effect in a held-out field the same chance of replicating, the share that replicated in the training fields. Field: forecasting. Excludes any use of an effect's own statistics.
- **logistic regression**. A model in which the log odds of success are a weighted sum of predictors, here the original z value, the log of the original sample size and the log of the replication sample size, fitted by maximum likelihood on the training fields. Field: statistics. Excludes any model of how originals were selected.
- **Brier score**. The mean squared difference between a forecast probability and the outcome coded 1 or 0, from 0 for a perfect forecast to 1 for a confident wrong one; 0.25 is what a forecast of one half always earns. Field: forecast verification. Excludes a ranking measure such as AUC, which ignores whether the probabilities are at the right level.
- **leave-one-field-out test**. A test in which each field in turn is held out, every parameter is fitted on the other fields, and forecasts are scored only on the held-out field. Field: statistics, cross-validation. Excludes scoring forecasts on the rows that were used to fit them.
- **95% interval**. The range that holds the middle 95 percent of the bootstrap or control draws for a number; we report it beside the number. Field: statistics. Excludes a statement that the true value lies inside with 95 percent probability.
- **project-cluster bootstrap**. An interval made by recomputing a statistic many times, here 2,000, on samples of whole replication projects drawn with replacement, so rows from one project stay together. Field: statistics. Excludes resampling single rows as if they were independent.

- **sign-flip permutation test**. A test that sums the per-project differences between two forecasts' scores, then flips the sign of each project's difference at random many times, here 4,000, to see how often a sum as large arises with no true difference. Field: statistics. Excludes a test that treats rows inside one project as independent.
- **Benjamini-Hochberg correction**. A rule that adjusts a set of p values so that, among the tests called significant, the expected share of false calls stays at or below a chosen rate q, here 0.05; the adjusted values are called q values. Field: statistics, multiple testing. Excludes the stricter Bonferroni rule, which bounds the chance of any false call.
- **BIC**. The Bayesian information criterion, minus twice the maximised log likelihood plus the number of parameters times the log of the number of rows; lower is better, and a difference of 10 is the usual bar for strong preference. Field: statistics, model comparison. Excludes a test of whether either model is right.
- **null distribution**. The spread of scores a control gives over many random draws; its 95th percentile is the bar a real score must clear. Field: statistics. Excludes a theoretical distribution assumed rather than drawn from the data.
- **calibration**. Agreement between forecast probabilities and observed frequencies: among effects given a chance near 0.7, about 70 percent should replicate; checked here in five bands of equal count. Field: forecast verification. Excludes discrimination, the separate question of whether higher forecasts go with more successes.
- **Verified line**. The closing statement of what was checked for publication, by what means, with what result and on what date; anything not checked is named as such. Field: our publication rule. Excludes a claim that the finding is true.
- **findings ledger**. The append-only, dated list of every result we have recorded, numbered F1, F2 and on; corrections are added as new dated lines and old entries are never rewritten. Field: our records. Excludes a curated list of successes.

15 References

[1] Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science 349(6251), aac4716. doi:10.1126/science.aac4716

[2] Open Science Collaboration. Reproducibility Project: Psychology, data files. Open Science Framework project ezcuj. CC0. Read 31 October 2026.

[3] FORRT Replication Database (FReD). Open Science Framework, doi:10.17605/OSF.IO/9r62x. CC BY 4.0. Read 31 October 2026.

[4] Röseler, L., Kaiser, L., Doetsch, C. and others. The Replication Database: documenting the replicability of psychological science. Open access at PubMed Central, PMC12270267.

[5] Gelman, A. and Carlin, J. (2014). Beyond power calculations: assessing Type S (sign) and Type M (magnitude) errors. Perspectives on Psychological Science 9(6), 641-651. doi:10.1177/1745691614551642

[6] Bartos, F. and Schimmack, U. (2022). Z-curve 2.0: estimating replication rates and discovery rates. Meta-Psychology 6, MP.2021.2720.

[7] Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., Kirchler, M., Nave, G. and Camerer, C. (2019). Predicting the replicability of social science lab experiments. PLoS ONE 14(12), e0225826. doi:10.1371/journal.pone.0225826