Investigating the replicability of the social and behavioural sciences

https://doi.org/10.1038/s41586-025-10078-y

Received: 3 March 2025

Accepted: 19 December 2025

Published online: 1 April 2026

Check for updates

Pursuing replicability — independent evidence for previous claims — is important for creating generalizable knowledge[1,2]. Here we attempted replications of 274 claims of positive results from 164 quantitative papers published from 2009 to 2018 in 54 journals in the social and behavioural sciences. Replications were high powered on average to detect the original effect size (median of 99.6%), used original materials when relevant and available, and were peer reviewed in advance through a standardized internal protocol. Replications showed statistically significant results in the original pattern for 151 of 274 claims (55.1% (95% confidence interval (CI) 49.2–60.9%)) and for 80.8 of 164 papers (49.3% (95% CI 43.8–54.7%)), weighed for replicating multiple claims per paper. We observed modest variation in replication rates across disciplines (42.5–63.1%), although some estimates had high uncertainty. The median Pearson's r effect size was 0.25 (95% CI 0.21–0.27) for original studies and 0.10 (95% CI 0.09–0.13) for replication studies, an 82.4% (95% CI 67.8–88.2%) reduction in shared variance. Thirteen methods for evaluating replication success provided estimates ranging from 28.6% to 74.8% (median of 49.3%). Some decline in effect size and significance is expected based on power to detect original effects and regression to the mean because we replicated only positive results. We observe that challenges for replicability extend across social–behavioural sciences, illustrating the importance of identifying conditions that promote or inhibit replicability[3,4].

A central aim of science is to discover regularities in nature. If a claimed discovery is true, independent researchers should be able to conduct a similar investigation and reach similar conclusions. A replication attempt involves testing the same research question as a previous investigation with independent evidence, whether the evidence is a new data collection or existing secondary data that were not used in the previous investigation[1,2].

Across several previous replication studies in the social and behavioural sciences, totalling hundreds of replication attempts, approximately half of well-powered replication studies provided statistically significant evidence in the same direction as an original finding[3,5–12]. Moreover, observed effect sizes in replication studies were about half as large as effect sizes in original studies on average[3]. Similar results have been observed in other fields, such as preclinical cancer biology[13].

Popular explanations for the low observed replication success rates include underreporting of negative or inconclusive evidence for claims; high sampling error from small samples and measurement error due to unreliable measures, coupled with a statistical threshold ($P < 0.05$) that serves as a publication filter; low rigour and quality control in research design and measurement; and questionable research practices that inflate the likelihood of obtaining positive outcomes[3,4,14–20]. These factors are compounded by a research culture that rewards novel and 'interesting' findings and discourages error correction[21–23]. Replication failures are not necessarily due to the original findings having low credibility. Low replication rates can also be due to false negatives, poorly designed replications, selecting only positive results for replication, and differences between original and replication studies that are initially perceived as unimportant[5,24]. More broadly, there are conceptual challenges in deciding what is a replication of a previous finding and how to assess whether a replication attempt succeeded. Attempting replications provides a grounded context to wrestle with those challenges.

Identifying the potential causes of replication failures assumes that the existing evidence of replication failures is itself replicable. Most evidence comes from studies that examine replication outcomes within a single discipline and with relatively small samples. Here we report a systematic replication of quantitative published claims conducted as part of the Defense Advanced Research Projects Agency (DARPA) Systematizing Confidence in Open Research and Evidence (SCORE) program. Papers and claims were drawn from a sample of well-known journals selected to represent a diversity of subdisciplines across the social and behavioural sciences (Supplementary Table 1). The selected journals were aggregated for expository purposes into six disciplines: 11 journals in business (including organizational behaviour, management and marketing), 9 journals in economics (including finance), 7 journals in education, 7 journals in political science (including public administration), 13 journals in psychology (including health) and 7 journals in sociology (including criminology). See Supplementary Information for outcomes separately by subdiscipline, selection effects, effect size comparisons, outcomes across claims and subset analyses (Extended Data Figs. 1–6 and Supplementary Tables 8–19). For more details on the topics of the replication studies, see Extended Data Figs. 7 and 8 and Supplementary Table 20. Papers and claims eligible for inclusion had to be quantitative research of any kind and contain a statistical inference that identified a positive result using non-simulated human data including any level of human organization (for example, individuals, families, political entities, firms and economic units).

A list of authors and their affiliations appears at the end of the paper.

Table 1 | Papers and claims selected and replication attempts by discipline

	Business	Economics	Education	Political science	Psychology	Sociology	Total
Stages of selecting claims and attempting replications							
Claims selected							
Papers with claims	766 (19.6)	673 (17.3)	445 (11.4)	551 (14.1)	950 (24.4)	515 (13.2)	3,900 (100)
Papers eligible for replication	294 (19.6)	255 (17.0)	172 (11.5)	212 (14.1)	369 (24.6)	198 (13.2)	1,500 (100)
Papers with multiple claims	38 (19.0)	33 (16.5)	23 (11.5)	32 (16.0)	49 (24.5)	25 (12.5)	200 (100)
Papers with single claim	256 (19.7)	222 (17.1)	149 (11.5)	180 (13.8)	320 (24.6)	173 (13.3)	1,300 (100)
Replications attempted							
Papers with replication started	46 (23.2)	27 (13.6)	14 (7.1)	18 (9.1)	65 (32.8)	28 (14.1)	198 (100)
Papers with replication attempts completed	36 (22.0)	24 (14.6)	13 (7.9)	15 (9.1)	58 (35.4)	18 (11.0)	164 (100)
Total replication attempts of claims[a]	42 (14.2)	40 (13.5)	28 (9.5)	45 (15.2)	108 (36.5)	33 (11.1)	296 (100)
Unique claims with replication attempts[b]	36 (13.1)	38 (13.9)	28 (10.2)	45 (16.4)	94 (34.3)	33 (12.0)	274 (100)

Data shown as *n* (%). [a]A count of all replication attempts with recognition that some claims were replicated multiple times (see Methods). [b]A count of how many claims had a replication attempt.

We first summarize how the sample of completed replications compares with the sample of papers and claims. Then, we assess the replication outcomes across various criteria[13]. Primary reporting emphasizes the two most reported replication metrics: statistical significance and effect size comparisons. This is followed by summarizing the same evidence with 12 additional metrics that have been used to evaluate replication success, each with different assumptions and limitations.

Replication attempts by discipline

We randomly selected papers and claims from those published within the selected journals and timeframe (see Methods). However, replication attempts were constrained by feasibility, access to resources and availability of researchers with relevant interest, expertise and instrumentation, leading to potential selection effects.

Table 1 illustrates the proportion of the selected papers by discipline for which replication attempts were started and completed. Representativeness across disciplines was maintained during identification and extraction of claims because we used random sampling strategies. Most of the change in representativeness occurred owing to the non-random process of selecting and starting a replication study: education and political science decreased in relative proportion of the sample, and psychology increased. Compared with the original sample of papers (first row), the proportion of completed attempts of unique claims (last row) was within 3.5% for economics, education, political science and sociology, and more notable variation was observed for business and psychology. Representativeness of replication attempts was relatively steady by year compared with the sample of papers as reported in the Supplementary Information.

We selected papers and attempted replications in two phases (see Methods), with 139 of the completed replications occurring from papers selected during the first phase and 25 from papers selected during the second phase. Replication success rates were similar between the first (49.5% statistically significant with the same pattern) and the second (48.0%) phase.

Outcomes by statistical significance

A common method for evaluating whether a replication supports an original claim is to assess whether the observed replication test statistic meets a statistical threshold (for example, $\alpha = 0.05$) with the effect showing the same pattern as the original evidence. This approach is simple to explain and can be applied to various statistical models. Yet, it also has some drawbacks, such as binary assessment and failure to incorporate indicators of precision[10,11,25–27].

For most papers, we attempted to replicate a single claim; for some papers, we attempted to replicate multiple claims. If the replication success of claims is dependent (perfectly correlated within study), then claims should be weighted so that each study counts as a single observation. If replication successes are independent, then each claim should be counted as a single observation. Weighting by paper, 80.8 of 164 replicated papers had statistically significant findings with the same pattern as the original finding (49.3% (95% CI 43.8–54.7%)), 16.0 had statistically significant findings with an opposing pattern (9.7% (95% CI 6.5–13.0%)) and 66.2 replications showed a null effect (40.4% (95% CI 34.9–45.8%)). Unweighted (by claim), 151 of 274 replicated claims had statistically significant findings with the same pattern (55.1% (95% CI 49.2–60.9%)), 24 had statistically significant findings with an opposing pattern (8.8% (95% CI 6.0–12.7%)) and 98 showed a null effect (35.8% (95% CI 30.3–41.6%)).

We only selected claims for replication that were identified as positive results, usually by exceeding a statistical threshold ($P < 0.05$). This tends to bias the sample against studies that underestimate effect sizes. We can estimate this expected regression to the mean by estimating the proportion of significant results that would have been selected without the requirement for statistically significant results. For a subsample of 200 papers, using the same process for identifying claims, we selected all outcomes from the paper, regardless of whether they were significant results[28]. We estimated that 2,747 of the 3,066 claims had significant results (89.6%). The high proportion of significant findings replicates a well-documented bias in the published literature favouring significance[18,19,29,30]. Nevertheless, because the significance rate was not 100%, some regression to the mean is expected in our replication success rates.

Replication attempts can also fail to produce a significant effect owing to limited statistical power. Statistical power varies depending on research designs and assumptions. Given the diversity of methods across studies, we calculated power for studies using two different approaches. Under the first approach, which was better suited for the original findings with standard effect sizes, the median power to detect the original effect size was approximately 99.6% ($\alpha = 0.05$): 90.2% of replications had at least 50% power to detect the original effect size, and 87.2% of replications had at least 75% power to detect the original effect size. In addition, under the approach used when a standard effect size could not be reliably approximated, the median power to detect the original effect size was 99.1% ($\alpha = 0.05$): 95.1% of replications had at least 50% power to detect the original effect size, and 84.5% of replications had at least 75% power to detect the original effect size. Under a combined approach in which each finding draws on the power approach most appropriate to it, the median power to detect the original effect size was 99.6% ($\alpha = 0.05$): 94.2% of replications had at least 50% power

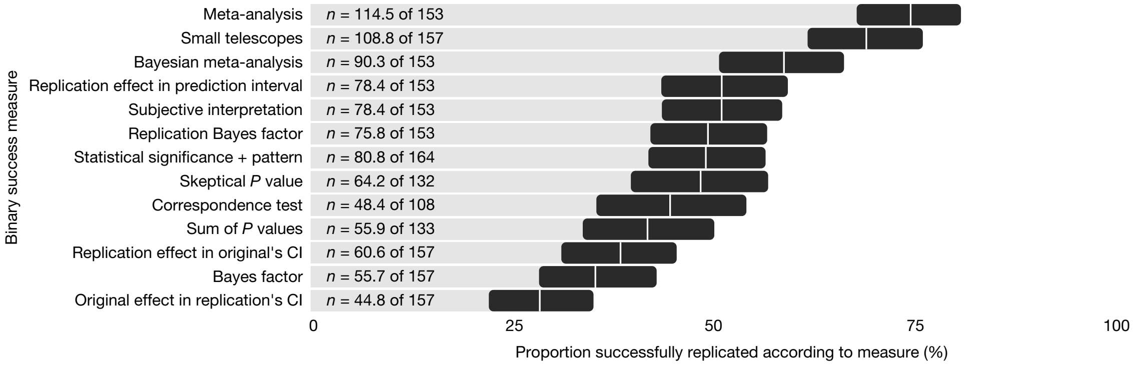


Fig. 1 | Replication success rates across 13 binary assessments for papers. The vertical white line in each row is the estimate of the percentage of papers replicated successfully, and the 95% confidence interval (CI) around the estimate is represented by the dark bar. The sample sizes are given as the weighted number of papers with successful replications followed by the number of papers to which that binary assessment could be applied. See Methods for explanations of each binary assessment.

to detect the original effect size, and 86.0% of replications had at least 75% power to detect the original effect size.

Outcomes by several binary assessments

Several methods have been proposed to assess replication success, each with advantages and disadvantages[31,32]. This project highlights a challenging problem: replication success metrics can be applied only to methods that meet their assumptions. In Fig. 1, we present 13 replication success metrics along with the number of papers to which each metric could be applied. Some are necessarily binary assessments of replication success or failure, and others were simplified to provide a binary assessment for comparison. The binary assessments were usable for 65.9–100.0% of the total sample papers and 55.8–100.0% of the total sample of claims. Only the statistical significance metric was applied in all cases. For some of those assessments, statistical assumptions needed to be applied to a subset of original and replication estimates.

Across metrics, replication success rates ranged from 28.6% to 74.8% with a median of 49.3%. The observed variation highlights the impact of the different assumptions underlying each approach. For example, the highest estimate of 74.8% for the meta-analytic combination of the original and replication evidence provides strong evidence of success because it includes original studies that provided evidence of success, and these tests are not independent of the original evidence.

Variation in the observed replication rates are affected by both the assumptions of the binary assessment and the subsample for which the metric could be used. Figure 2 presents correlations between each pair of binary assessments for replication outcomes to which both methods could be applied. Spearman correlations were positive and relatively high, with some exceptions (median 0.51, range 0.20–0.98. For example, analysts almost always interpreted success on the basis of statistical significance (as reflected by a correlation of 0.98), whereas measures that used confidence or prediction intervals showed relatively strong correlations with each other (0.60, 0.62 and 0.74) and weaker relations with other measures, partly because they treated replication outcomes both smaller and larger than original outcomes as failures to replicate (median 0.35, range 0.20–0.73).

Outcomes by effect size

Replication can also be defined as the correspondence between effect size coefficients observed in original studies and in their replications. Regardless of whether original studies and replication studies fall on the same side of a statistical threshold, they should produce effects of about the same magnitude. For the purposes of this project, we sought effect size metrics that would be as standardized as possible.

Figure 3 represents the replication effect sizes plotted against the original effect sizes for those claims that could be computed using the Pearson's *r* effect size. Original and replication effect sizes were positively correlated (Spearman's correlation of 0.43). The median effect size was 0.25 (95% CI 0.21–0.27) for original studies and 0.10 (95% CI 0.09–0.13) for replication studies, a 58.1% (95% CI 44.2–65.0%) reduction in correlation and a 82.4% (95% CI 67.8–88.2%) reduction in shared variance. Data points below the diagonal line are cases in which the replication effect size was smaller than the original effect size. The blue data points were statistically significant replications with the same pattern as the original; the red data points were not statistically significant or, if they were below 0, showed a different pattern than the original. Of 157 papers, 126.0 (80.3%) had a smaller effect size for the replications than the original studies, and 175 of 249 claims (70.3%) had a smaller effect size for the replications than the original studies. Table 2 provides summary statistics of effect sizes by papers and by claims.

Outcomes by discipline and year

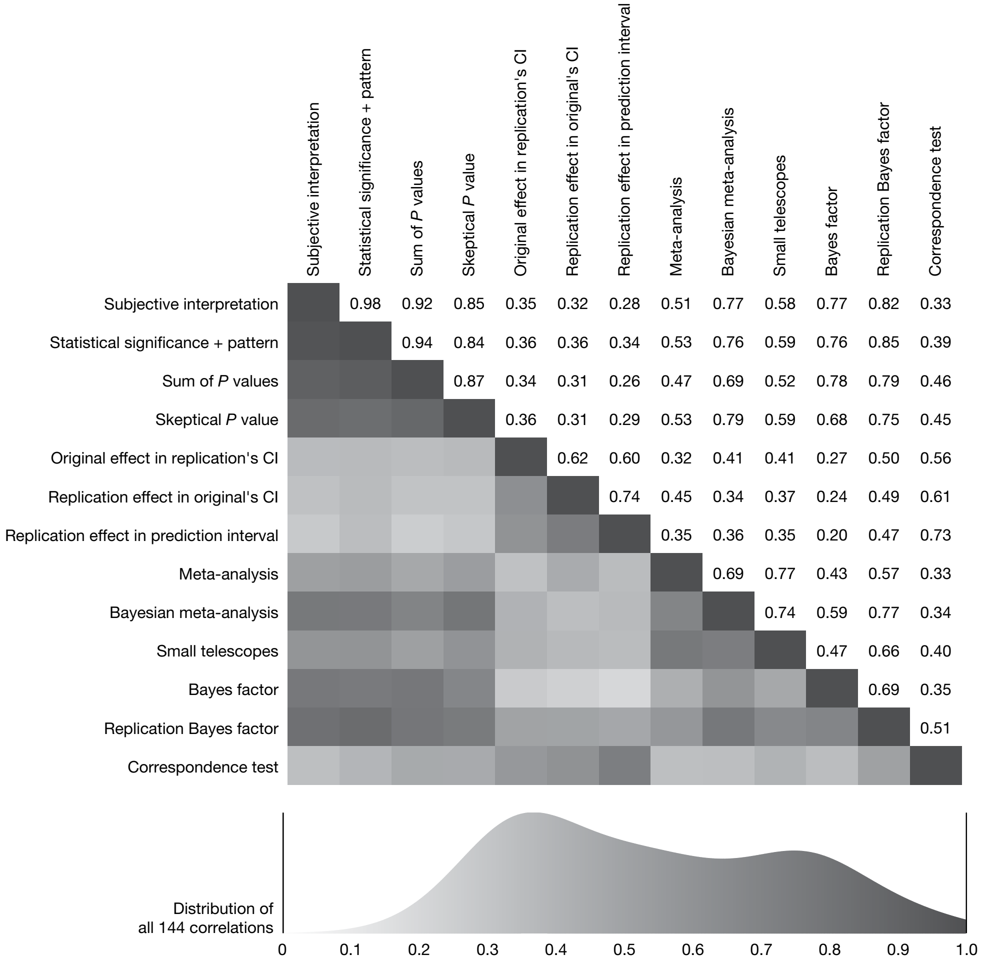
Table 3 summarizes original and replication outcomes separately by discipline for statistical significance and effect size. By discipline, based on statistical significance, successful replication rates ranged from 42.5% to 63.1% weighted across papers (median of 50.0%; χ^2 *P* = 0.76). A range of 45.5–71.4% was observed unweighted across claims (median of 50.8%; χ^2 *P* = 0.13). A similar analysis, reported in the Supplementary Information, examined variation of replication success by publication year across papers and did not show a significant effect by year (*P* = 0.14).

Outcomes by new or secondary data

Some replication attempts involved collecting new data, and others involved finding secondary data that were not used in the original research (Table 4). Replication attempts using new data were 0.930 (95% CI 0.651–1.338) times as likely as those using secondary data to have outcomes that were statistically significant and with the same pattern (unweighted by claims, they were 0.986 (95% CI 0.757–1.253) times as likely), suggesting similar replicability.

For outcomes that could be estimated with a Pearson's *r* effect size, replication attempts using new data produced effects that were less than half the size of their original effects across papers and about half across claims. Replication attempts using secondary data showed less decline than those using new data, but from original findings that had smaller effect sizes on average.

The effect size comparisons imply higher replicability for secondary versus new data replication attempts, in contrast to similar replicability observed on the statistical significance metric (Table 4). This could occur if the power of secondary data replication attempts was weaker.



	Subjective interpretation	Statistical significance + pattern	Sum of *P* values	Skeptical *P* value	Original effect in replication's CI	Replication effect in original's CI	Replication effect in prediction interval	Meta-analysis	Bayesian meta-analysis	Small telescopes	Bayes factor	Replication Bayes factor	Correspondence test
Subjective interpretation		0.98	0.92	0.85	0.35	0.32	0.28	0.51	0.77	0.58	0.77	0.82	0.33
Statistical significance + pattern			0.94	0.84	0.36	0.36	0.34	0.53	0.76	0.59	0.76	0.85	0.39
Sum of *P* values				0.87	0.34	0.31	0.26	0.47	0.69	0.52	0.78	0.79	0.46
Skeptical *P* value					0.36	0.31	0.29	0.53	0.79	0.59	0.68	0.75	0.45
Original effect in replication's CI						0.62	0.60	0.32	0.41	0.41	0.27	0.50	0.56
Replication effect in original's CI							0.74	0.45	0.34	0.37	0.24	0.49	0.61
Replication effect in prediction interval								0.35	0.36	0.35	0.20	0.47	0.73
Meta-analysis									0.69	0.77	0.43	0.57	0.33
Bayesian meta-analysis										0.74	0.59	0.77	0.34
Small telescopes											0.47	0.66	0.40
Bayes factor												0.69	0.35
Replication Bayes factor													0.51
Correspondence test													

Fig. 2 | Correlation matrix among binary assessments of replication success across papers. Correlation values are right of the diagonal, and the correlation magnitude is visualized left of the diagonal, with darker shading indicating stronger correlations.

Median power estimates across papers are slightly consistent with this possibility (secondary data 99.0%; new data 99.7%). Mean power estimates are more consistent because of a few more weakly powered secondary data replications (secondary data 86.8%; new data 97.3%), and by comparing median power to detect 75% of the original effect size (Table 5).

Figure 4 reproduces the scatterplots comparing original and replication effect sizes separately for new data (left) and secondary data (right). Among effect sizes that could be converted to Pearson's *r*, 86.8% (83.3 of 96) of new data replication attempts and 70.0% (42.7 of 61) of secondary data replication attempts had a weaker effect size than the original study. Note that disciplines differed markedly in the proportion of new versus secondary data replication attempts, with business and psychology being mostly new data replications, and education and sociology being mostly secondary data replications (see Supplementary Table 13).

Discussion

About half of the findings from a sample of social and behavioural science papers published from 2009 to 2018 replicated successfully with variation in success estimates across 13 binary assessments and effect size comparisons. Variation in replicability across the disciplines within the social and behavioural sciences was modest, with replication rates between 42.5% and 49% on the statistical significance metric for fields that had more than 20 replications. These findings are consistent with the cumulative evidence across systematic replications in the social and behavioural sciences[3,5,10] and from other fields[13], and they illustrate that there is substantial uncertainty in estimating replicability.

Assessing replicability

There are conceptual, methodological and inferential challenges to assessing replicability.

Conceptually, it can be challenging to attempt a replication of a previous finding. Strictly speaking, there is no such thing as exact replication. Replications inevitably differ in many ways including the units, treatments, observations and settings from the original research. Researchers must make decisions about how to conduct a good faith replication of an original claim. For example, should a present-day replication of a 2009 US study of political behaviour that used President Obama as a stimulus use Obama again, use the current US president or use the leader of the participants' nation? The answer depends on what features of that stimulus are essential for testing the original claim. The decision that a new study is a replication of a previous study is a theoretical commitment that they are testing the same claim[1]. The planning and review process of replication studies emphasized making design decisions that would produce a good-faith attempt to replicate the original finding. However, theoretical commitments can be wrong. When ostensible replications produce different results from original studies, it is common and reasonable for the subsequent debate to centre on whether it should be considered a replication. The methodology for the replicated studies is available in the Supplementary Information. In particular, differences and deviations identified by the replication team between the replication and the original study

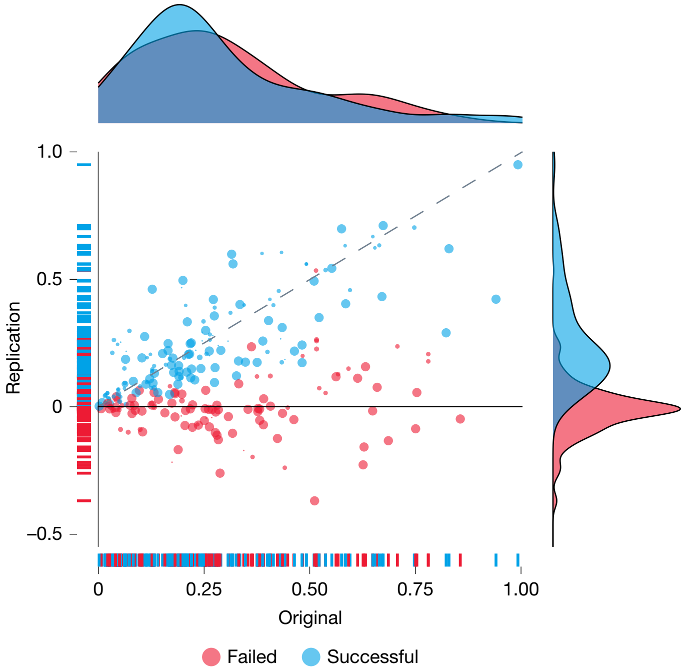


Fig. 3 | Scatterplot of Pearson's *r* effect sizes for original and replication studies. Each data point represents the estimated original and replication effect sizes for replicated claims. The size of the bullet is proportional to the number of claims there are per paper to illustrate paper weighting. Replication effect sizes are positive if the observed relationship has the same pattern as the original effect size, and negative if the observed relationship has a different pattern. Solid line indicates a zero replication effect size; dashed line indicates equivalent original and replication effect sizes. Data points are classified as successful for effect sizes that achieved statistical significance ($P < 0.05$, two-sided without adjustment for multiple comparisons) with the same pattern as the original study, and failed for effect sizes that did not (compare to Fig. 3 in ref. 5.

and preregistered replication design are extracted from individual reports and highlighted in Supplementary Tables 4–6.

Methodologically, it can be challenging to determine how to measure replication success. We used 13 binary metrics and compared effect sizes. Each approach has features that affect their usability across research designs and statistical models. No singular success metric has been accepted as optimal and universally applicable in the literature.

Inferentially, it can be challenging to determine whether original and replication studies produced the same outcomes. Each of the replication assessment criteria has strengths and weaknesses and may be based on different assumptions. For example, metrics that combine original and replication evidence are not independent tests of replicability and have relatively low sampling error. These tended to suggest the highest success rates. They might be useful only when it is safe to assume no selection or publication bias and the emphasis is on cumulative evidence. Conversely, metrics comparing original and replication effect sizes assume that replication effect sizes significantly smaller and larger than original effect sizes are failures to replicate. These tended to be among the lowest success rates; they might be mostly applicable under conditions in which the precise estimate is important versus knowing that the effect is larger than zero. Finally, a reviewer suggested that the metric relying on subjective assessment may be biased by the sample of researchers who participated in this replication project. The raters in this case almost perfectly mimicked using statistical significance for assessing replication success, but other raters might use different criteria. Reasonable minds may disagree on the best way to assess replication success. Productive follow-on investigations will use this dataset to further evaluate the merits of these metrics.

Table 2 | Original and replication findings by Pearson's *r* effect size by papers and claims

	Papers (weighted)		Claims (unweighted)	
	Original	Replication	Original	Replication
Outcomes				
Number of *r* effect sizes	157		249	
Median (IQR) sample size	206 (125.0)	545 (347.0)	236 (3,339.2)	556 (3,016.0)
Median Pearson's *r* effect size (s.d.)	0.25 (0.21)	0.10 (0.17)	0.24 (0.21)	0.13 (0.17)

The sample for this table is the studies for which a Pearson's *r* could be calculated for the original and replication outcomes. Papers are weighted combinations of claims accounting for multiple claims per paper replicated in some cases. IQR, interquartile range.

Table 3 | Original and replication outcomes by statistical significance and pattern and by Pearson's *r* effect size for six disciplines

Discipline	Replication attempt statistical significance and pattern		Median Pearson's *r* effect size (s.d.)	
	Counts	Percentage	Original estimate	Replication estimate
Business	17.0 of 36	47.2	0.24 (0.12)	0.10 (0.13)
Economics	10.2 of 24	42.5	0.28 (0.24)	0.13 (0.20)
Education	8.2 of 13	63.1	0.15 (0.28)	0.11 (0.10)
Political science	7.8 of 15	52.0	0.16 (0.22)	0.05 (0.16)
Psychology	28.4 of 58	49.0	0.29 (0.22)	0.11 (0.20)
Sociology	9.2 of 18	51.1	0.10 (0.16)	0.03 (0.17)

Papers are weighted combinations of claims accounting for multiple claims per paper replicated in some cases. The left column indicates the number of successful replications out of the total number of papers with replication attempts. Samples for the right columns are the papers for which a Pearson's *r* could be calculated for the original and replication outcomes.

In summary, the question 'did it replicate' can be difficult to answer. Fortunately, the answer for any given study does not matter much in the long run. Research is conducted on a study-by-study basis; replicability is established via a cumulative body of evidence. Over time, evidence accumulates and explanations mature. The explanations anticipate and account for variation across studies and contexts. The importance of deciding whether any two studies showed similar results fades away.

Understanding replicability

Failure to replicate does not mean the original claim was wrong.

A single failure to replicate does not justify concluding that the original research was wrong. Even if the replication was perfectly designed, the outcome could be missed or underestimated because of sampling error: a false negative. Even if the replication appeared to be testing the same research question, there could be differences in the methodology, sample or context that are unrecognized moderators of the outcome. In addition, even if the replication researchers were diligent in conducting the research, there could be unrecognized errors or flaws in implementing the replication protocol that interfered with observing the outcome.

We attempted to minimize these reasons for failing to replicate by using research designs that were well powered to detect the original effect size. We also obtained and adapted original materials whenever possible, conducted peer review in advance, preregistered the replication studies and promoted accountability by committing that materials and data available would be publicly accessible for review to the extent possible. These efforts provide some confidence in the rigour of the replication studies, but do not justify treating the outcomes as sacrosanct.

Table 4 | Original and replication outcomes by statistical significance and effect size by new or secondary data replications

	Papers		Claims	
	Original outcome	Replication outcome	Original outcome	Replication outcome
Outcomes				
Statistical significance and same pattern (%)				
New data replications	98 of 98 (100.0)	46.9 of 98 (47.8)	128 of 128 (100.0)	70 of 128 (54.7)
Secondary data replications	66 of 66 (100.0)	33.9 of 66 (51.4)	146 of 146 (100.0)	81 of 146 (55.5)
Median Pearson's *r* effect size (s.d.)				
New data replications	0.28 (0.18)	0.11 (0.19)	0.27 (0.18)	0.15 (0.19)
Secondary data replications	0.13 (0.23)	0.10 (0.14)	0.13 (0.23)	0.10 (0.15)

Original outcome refers to the published finding that was the target of the replication study. Replication outcome refers to the results of the replication attempt. New data replications are those that required data collection. Secondary data replications are those that used existing data that were independent of the original investigation. Papers are weighted combinations of claims accounting for multiple claims replicated in some papers.

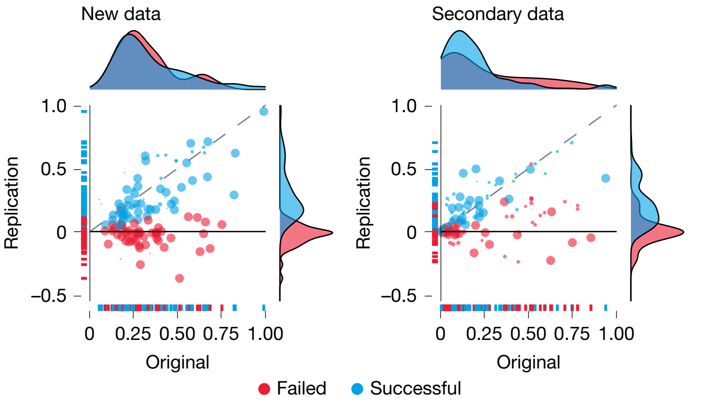


Fig. 4 | Scatterplot of Pearson's *r* effect sizes for original and replication outcomes for new data and secondary data replication attempts. New data (left) refers to replication attempts involving data collection ($n = 126$). Secondary data (right) refers to replication attempts based on existing data ($n = 123$). Each data point represents the estimated original and replication effect sizes for replicated claims. The size of the bullet is proportional to the number of claims there are per paper, to illustrate paper weighting. Replication effect sizes are positive if the observed relationship has the same pattern as the original effect size and negative if the observed relationship has a different pattern. Solid line indicates a zero replication effect size; dashed line indicates equivalent original and replication effect sizes. Data points are classified as successful for effect sizes that achieved statistical significance ($P < 0.05$, two-sided without adjustment for multiple comparisons) with the same pattern as the original study and failed for effect sizes that did not.

Successful replication does not mean the original claim was right. A single successful replication does not justify concluding that the original research was correct. The results of the original and replication studies could both be observed because of sampling error: a false positive. More importantly, the replicability of an effect is not the same as the validity of its interpretation. Original and replication studies may share confounds, faulty measures or other design weaknesses that produce replicable, but misinterpreted, outcomes.

The optimal replicability rate is not known. For example, in discovery contexts, it is understood that taking risks on unlikely possibilities will produce many false leads and occasional big rewards. Conducting replications helps to reveal weak spots and dead ends, identify boundary conditions, and mature theoretical predictions and explanations that improve replicability over time. In translation of research claims to policy and practice, it may be more important to have established high replicability to have confidence in their applicability and effectiveness.

The problem to solve is not unreplicability per se, it is overconfidence. Published and true are not synonyms[22], and the uncertainty of published claims may be underestimated. For many published findings, it is uncertain whether they will replicate at all, whether they are robust to minor variations in the research context, whether they are generalizable to other contexts, and whether they are valid interpretations of the evidence. Recognition of uncertainty will reduce overconfidence and increase recognition of the value of conducting replications and other verification methods to confront present understanding[1].

Constraints on generalizability

The sample for this research was a selective representation of the social–behavioural sciences. The inclusion criteria required a positive claim that is supported by a statistical inference. This was practically sensible for the purposes of the program, but it does not cover all relevant research. For example, we excluded research claiming a null result and qualitative research. Our results cannot be expected to generalize to these[33–35].

Table 5 | Median power to detect 75% of the original effect size

	Papers	Claims
Replication type		
New data replications (%)	94.5	94.1
Secondary data replications (%)	83.7	89.7

Original effect size refers to the published finding that was the target of the replication study. New data replications are those that require data collection. Secondary data replications are those that used existing data that were independent of the original investigation.

The sample covered a wide range of the social and behavioural sciences. However, the selection of relatively prominent journals might have led to higher or lower replicability than what would be observed if less-prominent journals were also included in the sample. Replicability might have been higher or lower if sampling had reached further back in history, and replicability might be changing in research published after the timeframe examined in this project.

Selection effects were minimized within the sample by using stratified random sampling of papers that met the inclusion criteria, but selection effects were introduced in attempting replications because some eligible papers were not matched with a replication team and some replication attempts were not completed (see Methods and Supplementary Information for more details). The main selection effect was the feasibility of conducting a replication given time and cost constraints. It is not difficult to generate plausible hypotheses that original findings from more resource-intensive research would be more, less or similarly replicable as original findings from less resource intensive research.

We did not explore correlations with replication outcomes that might help to advance understanding of the reasons for replication success and failure. An initial exploration using this dataset, in which modest correlations were observed between replication outcomes and several other potential indicators of research credibility, is reported by Abatayo and colleagues[28]. Many other variables could be investigated, such as risk of bias in research designs, sample sizes and evaluation of the impact of differences between original and replication studies.

We also presented outcomes from several different replicability metrics without evaluating their relative merits. A productive line of inquiry would interrogate the relationship between the underlying assumptions of the replicability measures and their effect on observed replicability rates. This will sharpen understanding of what a claim of replication success or failure means, and foster innovation or convergence on how to measure it. The dataset is openly available to stimulate further exploration.

Conclusion

The conditions that promote or inhibit replicability and how to assess it are worthy of additional investigation. Understanding the factors associated with the reliability of evidence will open pathways for advancing theory about research credibility and support pragmatic decision-making for translating research insights into practice[36].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-10078-y.

1. Nosek, B. A. & Errington, T. M. What is replication? *PLoS Biol*. **18**, e3000691 (2020).
2. *Reproducibility and Replicability in Science* (National Academies of Sciences, Engineering and Medicine, 2019).
3. Nosek, B. A. et al. Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol*. **73**, 719–748 (2022).
4. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav*. **1**, 0021 (2017).
5. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
6. Klein, R. A. et al. Many Labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci*. **1**, 443–490 (2018).
7. Klein, R. A. et al. Investigating variation in replicability: a "many labs" replication project. *Soc. Psychol*. **45**, 142–152 (2014).
8. Ebersole, C. R. et al. Many Labs 5: testing pre-data collection peer review as an intervention to increase replicability. Preprint at *OSF* https://doi.org/10.31234/osf.io/sxfm2 (2019).
9. Ebersole, C. R. et al. Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol*. **67**, 68–82 (2016).
10. Camerer, C. F. et al. Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
11. Camerer, C. F. et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat. Hum. Behav*. **2**, 637–644 (2018).
12. Cova, F. et al. Estimating the reproducibility of experimental philosophy. *Rev. Philos. Psychol*. **12**, 9–44 (2018).
13. Errington, T. M. et al. Investigating the replicability of preclinical cancer biology. *eLife* **10**, e71601 (2021).
14. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med*. **2**, e124 (2005).
15. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci*. **14**, 365–376 (2013).
16. Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. & Kievit, R. A. An agenda for purely confirmatory research. *Perspect. Psychol. Sci*. **7**, 632–638 (2012).
17. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci*. **22**, 1359–1366 (2011).
18. Greenwald, A. G. Consequences of prejudice against the null hypothesis. *Psychol. Bull*. **82**, 1–20 (1975).
19. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychol. Bull*. **86**, 638–641 (1979).
20. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606 (2018).
21. Giner-Sorolla, R. Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspect. Psychol. Sci*. **7**, 562–571 (2012).
22. Nosek, B. A., Spies, J. R. & Motyl, M. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci*. **7**, 615–631 (2012).
23. Nosek, B. A. et al. Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
24. Freese, J. & Peterson, D. Replication in social science. *Annu. Rev. Sociol* **43**, 147–165 (2017).
25. Andrews, I. & Kasy, M. Identification of and correction for publication bias. *Am. Econ. Rev*. **109**, 2766–2794 (2019).
26. Patil, P., Peng, R. D. & Leek, J. T. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci*. **11**, 539–544 (2016).
27. Valentine, J. C. et al. Replication in prevention science. *Prev. Sci*. **12**, 103–117 (2011).
28. Abatayo, A. L. et al. Credibility assessments in the social and behavioral sciences. Preprint at *MetaArXiv* https://doi.org/10.31222/osf.io/7u58q_v1 (2025).
29. Fanelli, D. "Positive" results increase down the hierarchy of the sciences. *PLoS ONE* **5**, e10068 (2010).
30. Fanelli, D. Negative results are disappearing from most disciplines and countries. *Scientometrics* **90**, 891–904 (2012).
31. Heyard, R. et al. A scoping review on metrics to quantify reproducibility: a multitude of questions leads to a multitude of metrics. *R. Soc. Open Sci*. **12**, 242076 (2025).
32. Muradchanian, J., Hoekstra, R., Kiers, H. & van Ravenzwaaij, D. How best to quantify replication success? A simulation study on the comparison of replication success metrics. *R. Soc. Open Sci*. **8**, 201697 (2021).
33. Peels, R. Replicability and replication in the humanities. *Res. Integr. Peer Rev*. **4**, 2 (2019).
34. Peels, R. & Bouter, L. The possibility and desirability of replication in the humanities. *Palgrave Commun*. **4**, 95 (2018).
35. TalkadSukumar, P. & Metoyer, R. Replication and transparency of qualitative research from a constructivist perspective. Preprint at *OSF* https://doi.org/10.31219/osf.io/6efvp (2019).
36. Hubbard, D. W. & Carriquiry, A. L. Quality control for scientific research: addressing reproducibility, responsiveness, and relevance. *Am. Stat*. **73**, 46–55 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Andrew H. Tyner[1], Anna Lou Abatayo[2], Mason Daley[1], Samuel Field[3], Nicholas Fox[1], Noah A. Haber[1], Krystal M. Hahn[1], Melissa Kline Struhl[4], Brinna Mawhinney[1], Olivia Miske[1], Priya Silverstein[5,6], Courtney K. Soderberg[1], Theresa Stankov[1], Ahmed Abbasi[7], Christopher L. Aberson[8], Balazs Aczel[9], Matúš Adamkovič[10,11,12], Nihan Albayrak[13,14], Peter J. Allen[15], Michael Andreychik[16], Eli Awtrey[17], Erick Axxe[18], Flavio Azevedo[19,20], Miles D. Bader[21], Bence Bago[22], James Bailey[23], Marjan Bakker[22], Gabriel Banik[24], George C. Banks[25], Ernest Baskin[26], Anatolia Batruch[27], Annika Beatteay[28], Sophie M. Behr[29,30], Nicholas Berente[7], Zachariah Berry[31], Jędrzej Białkowski[32], Bojana Bodroža[33], Laura Boeschoten[20], Miklos Bognar[9], Christian Bokhove[34], Diane Bonfiglio[35], Robin Bouwman[36], Timothy F. Brady[37], Scott R. Braithwaite[38], Gabriel Briceño Jiménez[39], Cameron Brick[40,41], Traci Bricka[42], Roman Briker[43], Annette N. Brown[44], Gordon D. A. Brown[45], Robbie C. M. van Aert[22], Kathryn Caldwell[46,47], Sara Capitan[48], Tabaré Capitán[48], Jesse Chandler[49], Tessa Charles[21], Christopher R. Chartier[35], Rahul Chawdhary[50], Kent Jason Cheng[51], William J. Chopik[52], Bruce Clark[53], Victoria E. Colvin[54], C. Cozette Comer[55], Giulio Costantini[56], Tom Coupé[32,57], Jamie Cummins[58], Aneta Czernatowicz-Kukuczka[59], Joshua de Leeuw[21], David Dobolyi[60], James N. Druckman[61], Jianhua Duan[32,62], Marin Dujmović[15], Daniel J. Dunleavy[63], Patrick K. Durkee[64,65], Cécile Emery[66], Kevin M. Esterling[67], Thomas R. Evans[68], Anna Fedor[69], Belén Fernández-Castilla[70], Nathan Fiala[71,72], James G. Field[73], Nathan Fong[74], Miguel A. Fonseca[66], Alexandra L. J. Freeman[19], Jeremy Freese[75], Sandra J. Geiger[76], Jing Geng[55], Laura M. Getz[77], Linda Marjoleine Geven[78], Ilka Helene Gleibs[79], Donna Pamella Gonzales[80], Janaki Gooty[25], Amélie Gourdon-Kanhukamwe[50,81], Cristina Greculescu[82,83], Siobhán M. Griffin[84], Lusine Grigoryan[85,86], Martina Grunow[87], Nicholas Gunby[57,88], Braeden Hall[53], Paul H. P. Hanel[89,90], Erin E. Hannon[91], Sam Harper[92], Marco Jürgen Held[93], Louis Hickman[55], Nathan C. Higgins[91,94], Svenja Hippel[95], Sven Hoeppner[12,96], Sanghyun Hong[32], Thomas J. Hostler[97], Michael Inzlicht[98,99], Kamil Izydorczak[100], Bastian Jaeger[22], Kristin Jankowsky[101], Johannes Jarke-Neuert[102,103], Matthew Jensen[104], Biljana Jokić[105,106], Daniel Jolles[14,89], Phillip Jolly[107], Angela M. Jones[108], Marie Juanchich[89], Pavol Kačmár[24], Hansika Kapoor[71,109], Andjela Keljanovic[33,110], Samjhana Koirala[71], Marta Kołczyńska[111], Dimitra Kouroupaki[112], Ulrich Kühnen[83], Michelangelo Landgrave[60,113], Michael J. Larson[38], Lyonel Laulié[39], Alice C. E. Lawrence[19], Joel M. Le Forestier[114], Katelin E. Leahy[52], Sungmok Lee[28], Jared Leslie[91], Savannah C. Lewis[35,115], Christopher Limnios[23], Hause Lin[4,116], An-Chiao Liu[20], John Wills Lloyd[117], Elliot A. Ludvig[45], Dermot Lynott[118], Jordan MacDonald[28], Peter Mallik[119], Daniel J. Mallinson[120], Daniele Marinazzo[121], Corinna S. Martarelli[122], Joshua Matacotta[123,124], Andrew McBride[125], Cillian McHugh[84], Gail McMillan[126], Esteban Méndez[127], Mitchell Metzger[35], Michalis P. Michaelides[128], Johannes Michalak[129], Leticia Micheli[78,130], Jeremy K. Miller[131], Marina Milyavskaya[126], Daniel C. Molden[132], Ambar G. Monjaras[91], David Moreau[133], Audrey Morrow[134], Cristóbal Moya[29,135], Liad Mudrik[136], Laetitia B. Mulder[137], Katie A. Munt[138], Arijit Nandi[92], Kathryn Nason[28], Carolin Nast[139], Gideon Nave[140], Heinrich H. Nax[141,142], Florian Neubauer[71,72], Phuong Linh L. Nguyen[143], Austin Lee Nichols[144], Gustav Nilsonne[145,146], Ernest O'Boyle[147], Jule Oettinghaus[86], Jeewon Oh[148], Adoril Oshana[25], Thomas Ostermann[129], Rachel P. Ostrowski[21], Abiola Oyebanjo[149], Radoslaw Panczak[58], Jamie Patrianakos[150], Ignacio Pavez[39,151], Yuri G. Pavlov[152], Sofia Persson[153], Marco Perugini[154], Kim Peters[66], Constant Pieters[155], Vladimir Ponizovskiy[86,156], Nathaniel D. Porter[55], Jason M. Prenoveau[157], Danka Purić[105], Mariah F. Purol[158], Arathy Puthillam[37,109], Kimberly A. Quinn[159], Marco Ramljak[20], W. Robert Reed[32,57], Michaela Ritchie[28], Margaret Ritzau[21], Sean Patrick Roche[108], Romina Rodela[160], Jan Philipp Röer[129], Ivan Ropovik[12,161,162], Jacob Rothschild[163], Justine Saal[86], Hani Safadi[164], Jason Samaha[134], Mary Sanchez[91], Soorya Sankaran[134], David Santos[165], Amanda C. Sargent[166], Marian Sauter[167], Kathleen Schmidt[35,53], Landon Schnabel[116], Amber N. Schroeder[42], Sebastian W. Schuetz[60], Brendan A. Schuetze[64,168], Michael Schulte-Mecklenbeck[58,169], Astrid Schütz[93], Eric L. Sevigny[170], Ellie Shackleton[84], Richard M. Shafranek[132], Samuel Shaki[171], Shishir Shakya[172], Miroslav Sirota[89], Matthew Ryan Sisco[173], Maksim M. Sitnikov[22], L. Robert Slevc[174], Laura Smalarz[175], Colin Tucker Smith[54], Joel S. Snyder[91], Nicolas Sommet[27], Fatih Sonmez[176], Barbara A. Spellman[117], Natalia Stanulewicz-Buckley[177], George Stock[25], Chris N. H. Street[178], Eirik Strømland[179], Tina Sundelin[145,146], Moin Syed[143], Anna Szabelska[180], Barnabas Szaszi[9,181], Ewa Szumowska[59,174], Anirudh Tagat[109], Susanne Täuber[40], Louis Tay[182], Stuti Thapa[183], Jason Thatcher[60,184], Domna Tsaklakidou[112], Lars Tummers[20], Elise Turkovich[134], Melba Verra Tutor[80], Karolina Urbanska[80], Anna Elisabeth van 't Veer[185], Marcel van Assen[20,22], Niels van de Ven[22], Ruben van den Goorbergh[20], Elisabeth Julie Vargo[6], Leigh Ann Vaughn[46], Simine Vazire[186], Jentien M. Vermeulen[187],

Diem Thi Hong Vo[57,188], Victor Volkman[71], Eric-Jan Wagenmakers[40], Deliah Wagner[189,190], Lukasz Walasek[45], Frank Walter[191], Lara Warmelink[192], Liuqing Wei[193], Marie Isabelle Weißflog[85,86], Nicholas Weller[67], Aaron L. Wichman[194], Jonathan Wilbiks[28], Jamal R. Williams[37], Kelly Wolfe[195], Finnian Wort[45], Ryan Wright[117], Jesper N. Wulff[196], Xindong Xue[197], Veronica X. Yan[64], Yuzhi Yang[28], Sangsuk Yoon[198], Iris Žeželj[105], Yinxian Zhang[199], Ignazio Ziano[200], Cristina Zogmaister[56], Zorana Zupan[105], Rolf A. Zwaan[36], Brian A. Nosek[1,117]✉ & Timothy M. Errington[1]

[1]Center for Open Science, Washington, DC, USA. [2]Wageningen University and Research, Wageningen, Netherlands. [3]SAS Institute, Cary, NC, USA. [4]Massachusetts Institute of Technology, Cambridge, MA, USA. [5]University of Coimbra, Coimbra, Portugal. [6]Institute for Globally Distributed Open Research and Education, Gothenburg, Sweden. [7]University of Notre Dame, Notre Dame, IN, USA. [8]The Dissertation Coach, Raleigh, NC, USA. [9]Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary. [10]Slovak Academy of Sciences, Bratislava, Slovakia. [11]University of Jyväskylä, Jyväskylä, Finland. [12]Charles University, Prague, Czech Republic. [13]Boğaziçi University, Istanbul, Turkey. [14]London School of Economics and Political Science, London, UK. [15]University of Bristol, Bristol, UK. [16]Fairfield University, Fairfield, CT, USA. [17]University of Cincinnati, Cincinnati, OH, USA. [18]Hendrix College, Conway, AR, USA. [19]University of Cambridge, Cambridge, UK. [20]Utrecht University, Utrecht, Netherlands. [21]Vassar College, Poughkeepsie, NY, USA. [22]Tilburg University, Tilburg, Netherlands. [23]Providence College, Providence, RI, USA. [24]Pavol Jozef Šafárik University in Košice, Košice, Slovakia. [25]University of North Carolina at Charlotte, Charlotte, NC, USA. [26]Saint Joseph's University, Philadelphia, PA, USA. [27]University of Lausanne, Lausanne, Switzerland. [28]University of New Brunswick, Saint John, New Brunswick, Canada. [29]DIW Berlin, Berlin, Germany. [30]Technical University of Berlin, Berlin, Germany. [31]University of Southern California, Los Angeles, CA, USA. [32]University of Canterbury, Christchurch, New Zealand. [33]Faculty of Philosophy, University of Novi Sad, Novi Sad, Serbia. [34]University of Southampton, Southampton, UK. [35]Ashland University, Ashland, OH, USA. [36]Erasmus University Rotterdam, Rotterdam, The Netherlands. [37]University of California San Diego, La Jolla, CA, USA. [38]Brigham Young University, Provo, UT, USA. [39]Facultad de Economía y Negocios, Universidad de Chile, Santiago, Chile. [40]University of Amsterdam, Amsterdam, Netherlands. [41]University of Inland Norway, Lillehammer, Norway. [42]University of Texas at Arlington, Arlington, TX, USA. [43]WHU – Otto Beisheim School of Management, Düsseldorf, Germany. [44]The Research and Evaluation Centre, Rockville, MD, USA. [45]University of Warwick, Coventry, UK. [46]Ithaca College, Ithaca, NY, USA. [47]The Analysis Factor, Ithaca, NY, USA. [48]Swedish University of Agricultural Sciences, Uppsala, Sweden. [49]University of Michigan, Ann Arbor, MI, USA. [50]Kingston University, London, UK. [51]University of Massachusetts Boston, University Park, PA, USA. [52]Michigan State University, East Lansing, MI, USA. [53]Southern Illinois University, Carbondale, IL, USA. [54]University of Florida, Gainesville, FL, USA. [55]Virginia Tech, Blacksburg, VA, USA. [56]Università degli Studi di Milano-Bicocca, Milan, Italy. [57]UCMeta, Christchurch, New Zealand. [58]University of Bern, Bern, Switzerland. [59]Jagiellonian University, Krakow, Poland. [60]University of Colorado Boulder, Boulder, CO, USA. [61]University of Rochester, Rochester, NY, USA. [62]Stats NZ, Christchurch, New Zealand. [63]Office of Program Policy Analysis and Government Accountability, Tallahassee, FL, USA. [64]University of Texas at Austin, Austin, TX, USA. [65]California State University, Fresno, Fresno, CA, USA. [66]University of Exeter, Exeter, UK. [67]University of California Riverside, Riverside, CA, USA. [68]University of Greenwich, London, UK. [69]Independent Researcher, Budapest, Hungary. [70]Universidad Nacional de Educación a Distancia, Madrid, Spain. [71]University of Connecticut, Storrs, CT, USA. [72]RWI - Leibniz Institute for Economic Research, Leibniz, Germany. [73]West Virginia University, Morgantown, WV, USA. [74]Rutgers University, Camden, NJ, USA. [75]Stanford University, Stanford, CA, USA. [76]University of Vienna, Vienna, Austria. [77]University of San Diego, San Diego, CA, USA. [78]Leiden University, Leiden, Netherlands. [79]London School of Economics, London, UK. [80]Independent Researcher, Quezon City, Philippines. [81]King's College London, London, UK. [82]Bremen International Graduate School of Social Sciences, Bremen, Germany. [83]Constructor University, Bremen, Germany. [84]University of Limerick, Limerick, Ireland. [85]University of York, York, UK. [86]Ruhr University, Bochum, Germany. [87]Leuphana University of Lüneburg, Lüneburg, Germany. [88]Contact Energy, Wellington, New Zealand. [89]University of Essex, Colchester, UK. [90]University of Bath, Bath, UK. [91]University of Nevada Las Vegas, Las Vegas, NV, USA. [92]McGill University, Montreal, Quebec, Canada. [93]University of Bamberg, Bamberg, Germany. [94]University of South Florida, Tampa, FL, USA. [95]University of Bonn, Bonn, Germany. [96]Freie Universität, Berlin, Germany. [97]Manchester Metropolitan University, Manchester, UK. [98]University of Toronto, Toronto, Ontario, Canada. [99]Rotman School of Management, Toronto, Ontario, Canada. [100]SWPS University, Warsaw, Poland. [101]University of Kassel, Kassel, Germany. [102]Forschungszentrum Jülich, Jülich, Germany. [103]Universität Hamburg, Hamburg, Germany. [104]University of Oklahoma, Norman, OK, USA. [105]University of Belgrade, Belgrade, Serbia. [106]Metropolitan University, Belgrade, Serbia. [107]Pennsylvania State University, University Park, PA, USA. [108]Texas State University, San Marcos, TX, USA. [109]Monk Prayogshala, Mumbai, India. [110]Faculty of Philosophy, University of Pristina-Kosovska Mitrovica, Kosovska Mitrovica, Serbia. [111]Institute of Political Studies of the Polish Academy of Sciences, Warszawa, Poland. [112]2nd Psychiatric Clinic, Atticon University General Hospital, Athens, Greece. [113]University of California, Riverside, Riverside, CA, USA. [114]University of Pittsburgh, Pittsburgh, PA, USA. [115]University of Alabama, Tuscaloosa, AL, USA. [116]Cornell University, Ithaca, NY, USA. [117]University of Virginia, Charlottesville, VA, USA. [118]Maynooth University, Maynooth, Ireland. [119]Hubbard Decision Research, Glen Ellyn, IL, USA. [120]Penn State Harrisburg, Middletown, PA, USA. [121]Ghent University, Ghent, Belgium. [122]UniDistance Suisse, Brig, Switzerland. [123]Western University of Health Sciences, Pomona, CA, USA. [124]Integrated Behavioral Health Research Institute, Pomona, CA, USA. [125]Santa Clara University, Santa Clara, CA, USA. [126]Carleton University, Ottawa, Ontario, Canada. [127]Central Bank of Costa Rica, San José, Costa Rica. [128]University of Cyprus, Nicosia, Cyprus. [129]Witten/Herdecke University, Witten, Germany. [130]Julius-Maximilian University of Würzburg, Würzburg, Germany. [131]Willamette University, Salem, OR, USA. [132]Northwestern University, Evanston, IL, USA. [133]University of Auckland, Auckland, New Zealand. [134]Univeristy of California, Santa Cruz, Santa Cruz, CA, USA. [135]Universität Bielefeld, Bielefeld, Germany. [136]Tel Aviv University, Tel Aviv, Israel. [137]University of Groningen, Groninge, Netherlands. [138]University of Queensland, Brisbane, Queensland, Australia. [139]UiS Business School, Stavanger, Norway. [140]Wharton School of Business, University of Pennsylvania, Philadelphia, PA, USA. [141]University of Zurich, Zurich, Switzerland. [142]ETH Zurich, Zurich, Switzerland. [143]University of Minnesota, Minneapolis, MN, USA. [144]Central European University, Vienna, Austria. [145]Karolinska Institutet, Stockholm, Sweden. [146]Stockholm University, Stockholm, Sweden. [147]Indiana University, Bloomington, IN, USA. [148]Syracuse University, Syracuse, NY, USA. [149]Policy Innovation Center, Abuja, Nigeria. [150]Loyola University Chicago, Chicago, IL, USA. [151]Business School, Universidad de San Andrés, Victoria, Argentina. [152]University of Tuebingen, Tuebingen, Germany. [153]Leeds Beckett University, Leeds, UK. [154]University of Milan-Bicocca, Milan, Italy. [155]Copenhagen Business School, Frederiksberg, Denmark. [156]Durham University, Durham, UK. [157]Loyola University Maryland, Baltimore, MD, USA. [158]Union College, Schenectady, NY, USA. [159]DePaul University, Chicago, IL, USA. [160]Södertörn University, Huddinge, Sweden. [161]Czech Academy of Sciences, Prague, Czech Republic. [162]University of Presov, Presov, Slovakia. [163]Verasight, Livingston, NJ, USA. [164]University of Georgia, Athens, GA, USA. [165]Universidad Autónoma de Madrid, Madrid, Spain. [166]Bentley University, Waltham, MA, USA. [167]Universität Ulm, Ulm, Germany. [168]University of Potsdam, Potsdam, Germany. [169]Max Planck Institute for Human Development, Berlin, Germany. [170]Georgia State University, Atlanta, GA, USA. [171]Ariel University, Ariel, Israel. [172]Appalachian State University, Boone, NC, USA. [173]Columbia University, New York, NY, USA. [174]University of Maryland, College Park, College Park, MD, USA. [175]Arizona State University, Glendale, AZ, USA. [176]Muş Alparslan University, Muş, Turkey. [177]Aston University, Birmingham, UK. [178]Keele University, Keele, UK. [179]Western Norway University of Applied Sciences, Bergen, Norway. [180]Psychological Science Accelerator, Ashland, OH, USA. [181]Corvinus Institute for Advanced Studies (CIAS), Corvinus University of Budapest, Budapest, Hungary. [182]Purdue University, West Lafayette, IN, USA. [183]University of Tulsa, Tulsa, OK, USA. [184]Temple University, Philadelphia, PA, USA. [185]Institute of Psychology, Leiden University, Leiden, Netherlands. [186]University of Melbourne, Parkville, Victoria, Australia. [187]Amsterdam UMC, Amsterdam, Netherlands. [188]RMIT University Vietnam, Ho Chi Minh City, Vietnam. [189]Center for Criminological Research Saxony (ZKFS), Chemnitz, Germany. [190]University of Jena, Jena, Germany. [191]Justus-Liebig-University Giessen, Giessen, Germany. [192]Lancaster University, Lancaster, UK. [193]Hubei University, Wuhan, China. [194]Western Kentucky University, Bowling Green, KY, USA. [195]Heriot-Watt University, Edinburgh, UK. [196]Aarhus University, Aarhus, Denmark. [197]Zhongnan University of Economics & Law, Wuhan, China. [198]University of Dayton, Dayton, OH, USA. [199]Queens College, CUNY, New York, NY, USA. [200]University of Geneva, Geneva, Switzerland. ✉e-mail: nosek@cos.io

Methods

This systematic replication effort was part of the SCORE program funded by DARPA to generate and evaluate automated measures of confidence in research claims[37]. Replications provided test data to evaluate the accuracy of human and machine predictions of replicability of claims. Evidence for reproducibility (same analysis, same data) and robustness (different analysis, same data) were also gathered as part of the programme. Relations among credibility assessments are reported in ref. 28. A full report of the SCORE methodology is accessible through this and several supporting papers[28,37,38]. Data, materials, code and other outputs from the programme are organized and publicly accessible for evaluation and re-use. This Methods section summarizes key features of sampling, conducting the replication studies, aggregating the data across replications and assessment of replication success.

Sampling and selecting claims

Claims to replicate were identified with a systematic selection process to reduce selection effects and increase generalizability of the findings to quantitative social and behavioural research. The project was conducted in two phases. The project started with a sample of 3,900 papers selected by a stratified random sampling from a larger set of papers to ensure representativeness across the 62 journals and publication dates from 2009 to 2018. From that pool, 600 papers were randomly selected during phase 1 as the papers eligible for conducting replication studies with a similar stratified random sampling process to maintain representativeness, and no additional random selection was conducted during phase 2 to constrain the sample of eligible papers ($n = 900$). This resulted in a total of 1,500 papers eligible for replication with 90.9% of claims subjected to replication attempts selected from the phase 1 portion. See ref. 28 for further details on the sampling frame and selection process.

Eligible papers were matched with research teams with relevant expertise to design and conduct the replication study. Here, random sampling is lost, because selection is based on feasibility, available resources and available expertise. Whenever possible, original methods and materials were collected from the original authors and adapted for the replication study. Replication teams prepared the research design, including the methodology and analysis plan, and put those through a peer review process that was managed by an independent editor and included independent reviewers plus at least one author of the original study if they agreed to provide review. Authors were instructed to design a 'good faith replication of the original claim', which could include keeping the methodology the same or updating it in service of improving the quality of the replication attempt. Peer reviewers and editors evaluated the replication design as a holistic assessment of how to improve it to be a good-faith attempt to replicate the original claim. For example, specific instructions for design and evaluation included statements such as "Remember that your goal in replication is to achieve a design that is a good faith test of the original claim (Nosek & Errington, 2020). Sometimes that is a straightforward repeat of the original procedure in your new sample. Sometimes that means adapting the methodology for the new context. Changing the methodology is not bad if it is done so in the service of improving the quality of the replication for testing the original claim". See the Supplementary Information for further details on the design and review process.

Replication designs could involve either the collection of new data or finding independent, existing data that were not used for the original research. Approved designs and analysis plans were preregistered on the Open Science Framework (OSF) before conducting the research. For the purposes of this project, initiating a draft of the preregistration for the replication study was the milestone defining that the replication had started.

In most cases, a single claim was identified in a single paper and subjected to a single replication attempt with independent data. Of the 1,500 papers eligible for replication, 1,300 had a single claim isolated for replication and 200 contained additional claims that could be replicated. For three claims, multiple replications were conducted using the same protocol, akin to 'many labs' studies[6,7,39]. In addition, for 15 claims, multiple replications were conducted using distinct protocols. For both of these cases, the primary reporting aggregates evidence across multiple replications of a single claim. Finally, there were 27 replications that added new data to data that had been used in the original research. These 'hybrid' replications were not included in the main text outcomes because the replications were not independent of the original studies, but they are reported in the Supplementary Information (Supplementary Tables 16–18).

Completed replication reports were reviewed for quality control by team members not involved in the replication study. Data, materials and code were archived on the OSF and made openly available to the maximum extent allowed without violating the privacy of participants or intellectual property licenses for any original materials. A total of 296 replications were conducted and, following aggregation evidence for multiple replications of a single claim, there were 274 replications of unique claims from 164 papers. See Extended Data Figs. 1–5, Supplementary Tables 1–3 and 7–9 and Supplementary Figs. 1–6 for details about sample selection, study design, attrition, statistical power and effect size estimation. See Extended Data Figs. 9 and 10 and Supplementary Table 19 for LLM-generated summaries of the topics and methods represented in the replicated papers.

Replication assessment metrics

We assessed the replicability of individual claims from papers that used diverse methodologies. We did this using statistical results from pairs of original and corresponding replication studies. In this section, we describe the approach for each of our 13 binary assessments of replication success.

Statistical significance and pattern. A common measure of concluding that there is evidence for an original claim and replication of that claim is achieving statistical significance ($P < 0.05$) with the hypothesized pattern of results. For papers to be included in the sample, the original research needed to have an outcome that could be assessed for replicability with this criterion.

Subjective interpretation. Subjective assessment of whether the original finding replicated successfully, provided by replication teams or project coordinators. No explicit constraints were provided to guide subjective interpretation, and the assessment may be contingent on the identities of the researchers making that subjective judgement. The only reason that this criterion was not used for some findings was because an interpretation was not collected or the interpretation was non-committal to being a success or failure.

Sum of *P* values. Calibrating the sum of original and replication *P* values can control the overall false-positive rate and enable replication success even if the original study was non-significant[40]. An unweighted sum of *P* values concludes that the replication succeeded if the sum of one-sided *P* values is less than 0.035 (or equivalently, if the sum of two-sided *P* values is less than 0.07). A weighted version can be used if there are concerns about the diagnosticity of the original evidence, such as the possibility of questionable research practices artificially reducing the *P* value. We used the unweighted version, because this method highlights the maximum success rate compared with downweighting the influence of the original study.

Sceptical *P* value. This criterion generates a prior using the data from the original result to construct a posterior with an associated credible interval that just overlaps with zero[41]. The sceptical *P* value assesses the extent to which the replication data are inconsistent with this sceptical prior. The logic is to define in advance how sceptical to be about the replication evidence to believe that the effect does not exist.

Replication confidence interval. This criterion assesses whether the original effect estimate was within the 95% confidence interval of the replication study. This assumes that the original effect was estimated without error and assesses whether it is different from the replication estimate. Replications can produce stronger, weaker or opposing effects than original studies and fail on this metric.

Original confidence interval. The complementary criterion is whether the replication effect estimate was within the 95% confidence interval of the original study. This assumes that the replication was estimated without error and assesses whether it is different from the original estimate. Replications can produce stronger, weaker or opposing effects than original studies and fail on this metric.

Replication in prediction interval. The 95% prediction interval has the same basic logic as the approaches using confidence intervals, except that it incorporates the precision of both the original and replication effect size in determining the boundaries. As such, this criterion is the most liberal of the interval-based methods, including considering some replications estimated near zero or with an opposing pattern to be successful.

Meta-analysis. The fixed-effect meta-analysis criterion combines original and replication evidence into a single estimate and assesses whether the combined evidence is statistically significant with the same pattern as the original study. Because all original studies were positive results, this criterion is necessarily generous to observing replication success, as it is not independent of the original evidence.

Bayesian meta-analysis. This criterion is the conceptual equivalent of meta-analysis in the Bayesian framework[42]. We used a fixed-effect model to quantify evidence of the effect being present versus absent across both studies. In our implementation, the outcome needed to have 'moderate,' 'strong' or 'extreme' evidence against the null to qualify as a success. The prior for the average effect size is centred at 0 with a standard deviation of 0.25. One thousand five hundred iterations per chain were used, with the log-marginal likelihood being estimated by numerical integration with a relative tolerance of 0.1.

Small telescopes. The small telescopes approach assesses whether replication results are consistent with an effect size that could have been detected in the original study[43]. This is calculated in two steps. First, compute the effect size that would have given the original study 33% power. Second, conduct a one-sided hypothesis test of whether the replication data can reject the null hypothesis that it is not smaller than that effect size. This approach recognizes the difficulty of providing evidence for the absence of an effect, so instead defines replication failure as demonstrating that the original study could not have provided evidence for an effect as small as was observed in the replication.

Bayes factor. The Jeffreys–Zellner–Siow Bayes factor is the conceptual equivalent of the standard null hypothesis significance test in the Bayesian framework[44]. It provides relative favourability for the null versus alternative hypothesis, indicating both the absence or the presence of an effect. In our implementation, the outcome needed to have a Bayes factor against the null of greater than 10 to qualify as a success, corresponding to the interpretation categories of 'strong,' 'very strong' or 'extreme' evidence.

Replication Bayes factor. Replication Bayes factor is an alternative to Jeffreys–Zellner–Siow Bayes factor that directly examines the replication evidence in comparison to the original study[44,45]. It provides relative evidence that the replication effect is similar to the original versus being absent. It can only be applied in cases of a non-zero result. In our implementation, the outcome needed to have a Bayes factor against the null less than 1 to qualify as a success.

Correspondence test. This criterion considers the correspondence in the effect size estimates between original and replication studies[46]. It combines comparing (1) whether the hypothesis that the effect sizes are the same can be rejected in terms of statistical significance with (2) an equivalence test evaluating the hypothesis that the observed difference in effect sizes is not larger than a predefined equivalence threshold. The correspondence test provides four outcomes: (1) equivalent, which is failing to reject that the effect sizes are the same and rejecting that the difference in effect sizes is larger than an equivalence threshold; (2) trivially different, which is rejecting that the effect sizes are the same and rejecting that the difference in effect sizes is larger than an equivalence threshold; (3) different, which is rejecting that the effect sizes are the same and failing to reject that the difference in effect sizes is larger than an equivalence threshold; and (4) indeterminate, which is failing to reject that the effect sizes are the same and failing to reject that the difference in effect sizes is larger than an equivalence threshold. There are enriched possibilities of considering these four outcomes independently using this dataset. For the purposes of creating binary outcomes for comparison with other approaches, we treated equivalent and trivially different as successful replications, different as failed replications, and left out indeterminate outcomes.

Data aggregation

For most studies, the original main finding and its corresponding replication findings used the same statistical methods and thus could be assessed on the same effect size scale. However, studies used different statistical methods, and thus also used different effect size scales (for example, Cohen's *d*, odds ratio and regression coefficient in a multilevel analysis, and so on). These measures cannot be meaningfully compared unless they are converted to a common scale.

We converted as many native effect sizes to partial correlation wherever possible to facilitate these comparisons. Most results could be converted using accepted formulae based on the *t*, *z* or *F* statistics. In the case of *t* statistics, the partial correlation is approximated by $t/\sqrt(t^2 + \text{residual degrees of freedom})$. For *z* statistics, it is approximated by $z/\sqrt(z^2 + N)$. And for *F* statistics, it is approximated by $\sqrt((F \times \text{numerator degrees of freedom})/(F \times \text{numerator degrees of freedom} + \text{denominator degrees of freedom}))$. Where appropriate, we implemented these using the effectsize R package[47]. Some analyses, such as multilevel regression or regressions with clustered standard errors, required a tailored approach to approximate the effective sample size or degrees of freedom for converting to standard effect sizes. In most instances of structural equation models, the standardized path coefficients are treated as proxies for the partial correlation, following convention. As needed, partial correlations were also approximated from chi-square statistics or from (log) odds ratios.

The procedures used for each conversion can be found in this OSF project (https://osf.io/uqegb/), where the names of each folder correspond to the study ID of the specific replication study. Files with an underscore of '_replication' feature the conversions for the replication findings, whereas files with an underscore of '_original' feature the conversions for the original findings replicated in that study.

Data analysis and inference

Statistics presented in the paper are largely in the form of descriptive statistics and precision estimates. Proportions of successful replications and similar statistics are aggregated to the paper level unless otherwise noted. Where there are multiple items per paper (for example, three claims assessed in replication attempts), the sub-level items (for example, claims assessed nested within papers) are weighted by simple proportion (for example, each of the three claims receives a weight of one-third). The code used to generate each statistic reported in this paper is provided in the data and code repositories.

All standard errors, confidence intervals and other metrics of statistical uncertainty are generated by simple clustered bootstrap. Statistical uncertainty for statistics aggregated to the paper level are clustered at the paper level using a clustered bootstrap procedure. Confidence intervals (95%) are estimated through percentile intervals of the bootstrapped sample distribution.

Inclusion and ethics

Researchers from 31 nations participated in designing, conducting and evaluating replications. Joining the collaboration was an open process, promoted via social media primarily by the Center for Open Science and the corresponding author. Various roles were defined to maximize opportunities for researchers with varying skills, areas of interest and access to resources to participate. Criteria for earning co-authorship were defined in advance so that researchers could make informed decisions about joining the collaboration. All replication studies reported in this article involved primary data collection from human participants (Ashland University 7-22-19-#091, 9-30-19-#105, 9-30-19-#106, 7-2-20#12, 1-31-20#8 and 9-8-21#40; Fairfield University 2712; Western Kentucky University 20-116 and 21-255; University of Nevada Las Vegas 1521828 and 1528491-4; University of San Diego 2020-70; Occidental College F19096; University of California, San Diego 191782SX; University of Exeter 003030, 001979, 003507, 004097, 004385, 004384, 488230 and 488231; London School of Economics and Political Science 1047; Southern Illinois University 200071 and 21097; University of Toronto 38581 and 38822; Northwestern University STU00211653 and STU00211686; Cornell University 1912009293, 2001009314, 2105010350, 2105010351 and 2109010548; University of Queensland 2020000052; Texas State University 7274 and 7336; Ruhr University Bochum 20-6866; University of Texas at Austin 2020-04-0114; Justus-Liebig-University Giessen 20-026-757; Ithaca College 89; Saint Joseph's University 1522885-1, 1548814-1, 1606422-1, 1606324-1, 1629093-1 and 1774195-1; Rochester Institute of Technology 2112119, 03042120, 02070620, 05041321, 02052721 and 01052721; University of Michigan HUM00173465; University of Pennsylvania 834860; University of California, Davis 1547826-1; University of North Carolina, Charlotte 19-0406, 19-0802 and 22-0005; University of Maryland 1542892-1; Pennsylvania State University STUDY00013895 and STUDY00018137; Vassar College 01.17.20.01; University of Dayton; Reed College 2020-S05-FF2 and 2020-S30-FF3; University of Texas at Arlington 2020-0151; University of Groningen RDMPFEB-20200109-10402; Purdue University 2020-14; Arizona State University STUDY00011369; University of Wyoming 20200113TC02627; Ariel University AU-SOC-SS-20200122; University of Milan-Bicocca RM-2020-234; University of New Brunswick 004-2020, 021-2021 and 017-2021; Brigham Young University 2020-024; University of California, Santa Cruz 3600; Montclair State University FY19-20-1652; Carleton University 112136; Jagiellonian University in Krakow 1556167-1; Loyola University Chicago 2908/6640; University of California, Riverside HS-20-003; Virginia Tech 20-027; Heinrich Heine University Dusseldorf 2020-766; University of Cambridge PRE.2020.011 and PRE.2020.086; Rutgers University-Camden Pro2019002539; University of Chile; Attikon General University Hospital 136/11-3-2020, 370/7-7-2020 and 376/19-7-2021; University of Minnesota STUDY00009691; University of Connecticut X20-0102, X21-0162 and H21-0079; North Dakota State University SM20283; Virginia Polytechnic Institute and State University 20-518; BRANY SBER IRB 20-041-771, 20-037-770, 20-042-772, 20-032-764, 20-072-839, 20-025-737 and 21-066-895) or used secondary analysis of data of organizations, firms or human participants (University of North Carolina, Charlotte 19-0804; BRANY SBER IRB 20-019-749 and 21-056-749). All replication studies underwent local ethics review to confirm that the research was performed in accordance with all relevant guidelines and regulations, and that informed consent was obtained where necessary. All protocols received concurrence from MRDC HRPO and NIWC-PAC HRPO.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data, materials and code associated with this research that can be shared without restriction are publicly available in a living OSF repository (https://doi.org/10.17605/OSF.IO/G5SNY)[48]. The living OSF repository represents improvements, fixes and additions that occur post-publication. Readers can also access a registered, archived version of this repository that is precisely the data, code and documentation as they existed upon publication of this paper (https://doi.org/10.17605/OSF.IO/BZFGY). The repository includes all available documentation for replication attempts regardless of whether they were completed. This includes most of the data and code from the individual replication attempts, save for any data that is proprietary or protected that will not be made available, or for which analyst teams were uncertain or unable to confirm that they were allowed to share secondary data. It is possible that some data, materials or code that could be shared openly is not available at the time of publication. Readers are encouraged to contact the corresponding author or the authors of the relevant sub-project (Supplementary Table 3) to see if more research content can be shared in the living repository. This paper is part of a collection of papers reporting on the SCORE program. Documentation, data and code for the entire program are available at https://doi.org/10.17605/OSF.IO/DTZX4.

Code availability

Code for individual replication projects is available alongside data and materials for each project in the OSF repository (https://doi.org/10.17605/OSF.IO/G5SNY). This includes a push button package with all code and data used to produce all statistics, figures and tables, and code that populates them directly into the manuscript from a template. Also available is a registered, archived version of the repository containing precisely the data, code and documentation used to generate the outcomes reported in this paper (https://doi.org/10.17605/OSF.IO/BZFGY).

37. Alipourfard, N. et al. Systematizing Confidence in Open Research and Evidence (SCORE). Preprint at *SocArXiv* https://doi.org/10.31235/osf.io/46mnb (2021).
38. Miske, O. et al. Investigating the reproducibility of the social and behavioral sciences. *Nature* https://doi.org/10.1038/s41586-026-10203-5 (2025).
39. Ebersole, C. R. et al. Many Labs 5: testing pre-data-collection peer review as an intervention to increase replicability. *Adv. Methods Pract. Psychol. Sci.* **3**, 309–331 (2020).
40. Held, L., Pawel, S. & Micheloud, C. The assessment of replicability using the sum of p-values. *R. Soc. Open Sci.* **11**, 240149 (2024).
41. Micheloud, C., Balabdaoui, F. & Held, L. Assessing replicability with the sceptical p-value: type-I error control and sample size planning. *Stat. Neerlandica* https://doi.org/10.1111/stan.12312 (2023).
42. Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M. & Wagenmakers, E.-J. A primer on Bayesian model-averaged meta-analysis. *Adv. Methods Pract. Psychol. Sci.* https://doi.org/10.1177/25152459211031256 (2021).
43. Simonsohn, U. Small telescopes: detectability and the evaluation of replication results. *Psychol. Sci.* **26**, 559–569 (2015).
44. Verhagen, J. & Wagenmakers, E.-J. Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* **143**, 1457–1475 (2014).
45. Ly, A., Etz, A., Marsman, M. & Wagenmakers, E.-J. Replication Bayes factors from evidence updating. *Behav. Res. Methods* **51**, 2498–2508 (2019).
46. Steiner, P. M., Sheehan, P. & Wong, V. C. Correspondence measures for assessing replication success. *Psychol. Methods* https://doi.org/10.1037/met0000597 (2023).
47. Ben-Shachar, M., Lüdecke, D. & Makowski, D. effectsize: Estimation of effect size indices and standardized parameters. *J. Open Source Softw.* **5**, 2815 (2020).
48. Haber N. et al. Materials for "Investigating the replicability of the social and behavioral sciences". *OSF* https://doi.org/10.17605/OSF.IO/G5SNY (2025).

Acknowledgements This work was supported by the DARPA under cooperative agreements no. N660011924015 (to principal investigator B.A.N.) and HR00112020015 (to principal investigator T.M.E.). The views, opinions, findings and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the US Government. We thank B. Arendt, A. Denis, M. Dirzo, Z. Loomas, B. Luis, L. Markham, E. S. Parsons and A. Russell for their contributions to this project.

Author contributions A.H.T., S.F., M.K.S., F.A., W.J.C., N. Fiala, N. Fox, A.L.J.F., K.E.L., S.L., F.N., R.R., J. Samaha, S. Sankaran, A.N.S., D.T., D.W., B.A.N. and T.M.E. conceptualized the study. A.H.T., N.A.H. and J. MacDonald curated the data. A.H.T., N.A.H., M.K.S., C.L.A., M.A., N.A., P.J.A., M. Andreychik, E. Axxe, F.A., B. Bago, J. Bailey, M. Bakker, G.C.B., E.B., A. Batruch, A. Beatteay, Z.B., M. Bognar, C. Bokhove, R. Bouwman, T.F.B., G.B.J., T.B., R. Briker, G.D.A.B., R.C.M.v.A., K.C., T. Capitán, J. Chandler, T. Charles, C.R.C., K.J.C., W.J.C., V.E.C., G.C., T. Coupé, J. Cummins, A.C.-K., J.d.L., D.D., J.N.D., J.D., M. Dujmović, P.K.D., C.E., T.R.E., A.F., B.F.-C., N. Fiala, J.G.F., L.M.G., I.H.G., J. Gooty, S.M.G., N.G., B.H., P.H.P.H., E.E.H., S. Harper, M.J.H., N.C.H., S. Hippel, S. Hoeppner, S. Hong, T.J.H., B. Jaeger, K.J., J.J.-N., A.M.J., M. Juanchich, P.K., H.K., S.K., M.K., D.K., M.L., L.L., A.C.E.L., J.M.L.F., K.E.L., J.L., S.C.L., C.L., A.-C.L., J. MacDonald, P.M., D.J.M., A. McBride, C. McHugh, G. McMillan, E.M., M.P.M., L. Micheli, M. Milyavskaya, D.C.M., A.G.M., A. Morrow, C. Moya, L.B.M., K.A.M., A.N., H.H.N., F.N., P.L.L.N., J.Oh, A. Oshana, A. Oyebanjo, R. Panczak, J.P., Y.G.P., S.P., M.P., K.P., C.P., V.P., N.D.P., D.P., M.F.P., A.P., M. Ramljak, W.R.R., S.P.R., J.P.R., I.R., J. Rothschild, J. Samaha, S. Sankaran, D.S., A.C.S., K.S., A.N.S., B.A. Schuetze, M.S.-M., E.L.S., S. Shaki, S. Shakya, M. Sirota, M.R.S., M.M.S., L.R.S., L. Smalarz, J.S.S., N.S., F.S., C.N.H.S., M. Syed, A. Szabelska, E. Szumowska, A.T., S. Täuber, L. Tay, S. Thapa, L. Tummers, E.T., M.v.A., N.v.d.V., E.J.V., L.A.V., D.T.H.V., V.V., D.W., L. Walasek, N.W., A.L.W., J.W., J.R.W., K.W., R.W., J.N.W., V.X.Y., S.Y., I. Žeželj and Y.Z. conducted formal analysis. A. Abbasi, F.A., G.B.J., D.D., A.L.J.F., J. Gooty, J.P.R., J. Samaha, A.C.S., A.N.S., B.A.N. and T.M.E. acquired funding. A.H.T., A.L.A., M. Daley, S.F., N. Fox, K.M.H., M.K.S., B.M., O.M., C.K.S., A.A., B.A., M.A., N.A., P.J.A., M. Andreychik, E. Axxe, F.A., M.D.B., B. Bago, J. Bailey, G.B., G.C.B., E.B., A. Batruch, A. Beatteay, S.M.B., N.B., Z.B., J.B., B. Bodroža, M. Bognar, C. Bokhove, D.B., R. Bouwman, T.F.B., S.R.B., G.B.J., T.B., R. Briker, G.D.A.B., K.C., S.C., T. Capitán, J. Chandler, T. Charles, C.R.C., K.J.C., W.J.C., V.E.C., C.C.C., G.C., T. Coupé, J. Cummins, A.C.-K., J.d.L., D.D., J.N.D., J.D., M. Dujmović, C.E., T.R.E., N. Fiala, J.G.F., N. Fong, M.A.F., A.L.J.F., J.F., J. Geng, L.M.G., I.H.G., D.P.G., J. Gooty, C.G., S.M.G., L.G., M.G., B.H., P.H.P.H., E.E.H., S. Harper, M.J.H., L.H., N.C.H., S. Hippel, S. Hoeppner, S. Hong, T.J.H., B. Jaeger, K.J., J.J.-N., M. Jensen, B. Jokić, D.J., A.M.J., M. Juanchich, A. Keljanovic, S.K., M.K., U.K., M.L., M.J.L., L.L., A.C.E.L., J.M.L.F., K.E.L., S.L., J.L., S.C.L., C.L., H.L., A.-C.L., E.A.L., J. MacDonald, P.M., D.J.M., A. McBride, C. McHugh, G. McMillan, M. Metzger, M.P.M., J. Michalak, L. Micheli, M. Milyavskaya, D.C.M., A.G.M., D. Moreau, A. Morrow, L.B.M., K.A.M., K.N., C.N., G. Nave, H.H.N., F.N., G. Nilsonne, E.O., J. Oettinghaus, J. Oh, A. Oshana, T.O., R.P.O., J.P., I.P., Y.G.P., S.P., M.P., K.P., V.P., N.D.P., D.P., M.F.P., A.P., M. Ramljak, W.R.R., M. Ritchie, S.P.R., J.P.R., I.R., J. Rothschild, J. Saal, H.S., J. Samaha, M. Sanchez, S. Sankaran, A.C.S., K.S., L. Schnabel, A.N.S., S.W.S., B.A. Schuetze, A. Schütz, E.L.S., E. Shackleton, R.M.S., S. Shaki, S. Shakya, M. Sirota, L.R.S., L. Smalarz, C.T.S., J.S.S., N.S., F.S., G.S., M. Syed, A. Szabelska, B.S., E. Szumowska, A.T., S. Täuber, L. Tay, J.T., L. Tummers, E.T., M.V.T., K.U., N.v.d.V., R.v.d.G., E.J.V., L.A.V., J.M.V., D.T.H.V., V.V., D.W., L. Walasek, F. Walter, L. Warmelink, L. Wei, M.I.W., N.W., A.L.W., J.W., J.R.W., K.W., V.X.Y., Y.Y., S.Y., I. Žeželj, Y.Z. and T.M.E. performed the investigation. A.H.T., A.L.A., S.F., N. Fox, N.A.H., M.K.S., B.M., O.M., P.S., C.K.S., M.A., N.A., P.J.A., M. Andreychik, E. Axxe, F.A., M.D.B., B. Bago, J. Bailey, G.C.B., E.B., A. Batruch, A. Beatteay, S.M.B., Z.B., B. Bodroža, M. Bognar, C. Bokhove, R. Bouwman, T.F.B., G.B.J., T.B., R. Briker, G.D.A.B., K.C., S.C., T. Capitán, J. Chandler, T. Charles, C.R.C., W.J.C., G.C., T. Coupé, J. Cummins, A.C.-K., J.d.L., P.K.D., C.E., T.R.E., J.G.F., J. Geng, L.M.G., I.H.G., J. Gooty, C.G., S.M.G., L.G., B.H., P.H.P.H., S. Harper, M.J.H., L.H., N.C.H., T.J.H. K.J., J.J.-N., B. Jokić, D.J., P.J., A.M.J., M. Juanchich, S.K., L.L., J.M.L.F., K.E.L., S.C.L., C.L., H.L., A.-C.L., E.A.L., J. MacDonald, D.J.M., C. McHugh, G. McMillan, M.P.M., L. Micheli, M. Milyavskaya, D.C.M., C. Moya, L.B.M., A.N., A.L.N., J.Oh, R.P.O., J.P., I.P., Y.G.P., S.P., M.P., K.P., V.P., N.D.P., D.P., M. Ramljak, M. Ritzau, S.P.R., R.R., J.P.R., I.R., J. Saal, J. Samaha, M. Sanchez, S. Sankaran, K.S., L. Schnabel, A.N.S., B.A. Schuetze, M.S.-M., E.L.S., R.M.S., S. Shaki, S. Shakya, M. Sirota, L.R.S., L. Smalarz, J.S.S., N.S., F.S., M. Syed, A. Szabelska, E. Szumowska, A.T., S. Täuber, L. Tay, D.T., K.U., M.v.A., N.v.d.V., E.J.V., L.A.V., S.V., D.T.H.V., D.W., L. Walasek, A.L.W., J.W., J.R.W., K.W., F. Wort, V.X.Y., S.Y., I. Žeželj, I. Ziano, B.A.N. and T.M.E. formulated the methodology. A.H.T., O.M., A.C.E.L., B.A.N. and T.M.E. provided project admin. A.H.T., N.A.H., A.L.A., M.K.S., T. Stankov, N.A., M. Andreychik, E. Axxe, F.A., M.D.B., B. Bago, J. Bailey, E.B., A. Batruch, A. Beatteay, S.M.B., Z.B., M. Bognar, C. Bokhove, R. Bouwman, T.F.B., R. Briker, R.C.M.v.A., K.C., T. Capitán, J. Chandler, T. Charles, C.R.C., K.J.C., V.E.C., G.C., T. Coupé, J. Cummins, J.d.L., P.K.D., T.R.E., B.F.-C., J.G.F., J.Geng, I.H.G., S.M.G., P.H.P.H., S. Harper, M.J.H., N.C.H., S. Hippel, S. Hoeppner, T.J.H., K.J., J.J.-N., D.J., A.M.J., M. Juanchich, P.K., S.K., L.L., S.C.L., C.L., H.L., A.-C.L., J. MacDonald, D.J.M., A. McBride, C. McHugh, L. Micheli, D.C.M., A. Morrow, C. Moya, A. Oshana, R.P.O., J.P., Y.G.P., K.P., C.P., V.P., N.D.P., M.F.P., A.P., M. Ramljak, S.P.R., J.P.R., I.R., J. Samaha, M. Sanchez, S. Sankaran, D.S., K.S., L. Schnabel, B.A. Schuetze, E.L.S., S. Shaki, S. Shakya, M. Sirota, M.R.S., L.R.S., J.S.S., N.S., F.S., C.N.H.S., M. Syed, A. Szabelska, E. Szumowska, A.T., L.Tay, E.J.V., L.A.V., D.T.H.V., V.V., D.W., L. Walasek, M.I.W., A.L.W., J.W., J.N.W., V.X.Y., S.Y. and Y.Z. provided software. A.H.T., K.M.H., M.K.S., O.M., C.K.S., F.A., G.C.B., E.B., A.N.B., R.C.M.v.A., T. Capitán, C.R.C., W.J.C., K.M.E., J. Gooty, A.G.-K., L.G., E.E.H., M.I., K.E.L., J.W.L., M. Milyavskaya, A. Morrow, G. Nilsonne, V.P., N.D.P., K.A.Q., W.R.R., R.R., J.P.R., J. Samaha, A.N.S., A. Schütz, E.L.S., S. Shaki, C.T.S., J.S.S., B.A. Spellman, D.T., E.-J.W., A.L.W., C.Z., B.A.N. and T.M.E. provided supervision. A.H.T., A.L.A., N.A.H., O.M., P.S., N.A., E. Awtrey, F.A., J. Bailey, G.C.B., E.B., J.B., L.B., S.R.B., G.B.J., C. Brick, A.N.B., T. Capitán, R.C., W.J.C., J. Cummins, M. Dujmović, D.J.D., K.M.E., B.F.-C., N. Fiala, S.J.G., L.M. Geven, I.H.G., A.G.-K., K.I., B. Jokić, P.J., H.K., K.E.L., C.L., J.W.L., J. MacDonald, D.J.M., D. Marinazzo, C.S.M., J. Matacotta, J.K.M., D. Moreau, A. Morrow, L. Mudrik, H.H.N., A.L.N., G. Nilsonne, J.Oh, A. Oshana, N.D.P., J.M.P., M.F.P., K.A.Q., J.P.R., H.S., S. Sankaran, D.S., M. Sauter, K.S., L. Schnabel, A.N.S., E.L.S., R.M.S., S. Shaki, S. Shakya, L. Smalarz, B.A. Spellman, N.S.-B., E. Strømland, T. Sundelin, A. Szabelska, L. Tummers, K.U., A.E.v.'t.V., N.v.d.V., E.J.V., V.V., E.-J.W., D.W., A.L.W., J.W., X.X., I. Ziano, C.Z., Z.Z., R.A.Z. and T.M.E. performed validation. A.H.T., N.A.H., B.M., F.A., G.B.J., M.Dujmović, N.C.H., J.MacDonald, J. Matacotta, A. Morrow, A. Oshana, M. Ramljak, J. Samaha, S. Sankaran, S. Shaki, D.W., A.L.W. and B.A.N. conducted visualization. A.H.T., B.M., O.M., F.A., G.B.J., W.J.C., B.C., C.C.C., A.F., B.F.-C., N. Fiala, S.K., K.E.L., S.L., J. MacDonald, J. Matacotta, A. Morrow, C. Moya, H.H.N., F.N., J. Oh, A. Oshana, M.F.P., M. Ritchie, R.R., D.S., A.C.S., A.N.S., S. Shaki, S. Shakya, N.S., S. Täuber, E.J.V., V.V., D.W., N.W., B.A.N. and T.M.E. wrote the original draft of the manuscript. A.H.T, M. Daley, N.A.H., B.M., O.M., P.S., M.A., N.A., F.A., M. Bakker, G.B., G.C.B., B. Bodroža, G.B.J., C. Brick, R. Briker, R.C.M.v.A., J. Chandler, C.R.C., W.J.C., B.C., C.C.C., A.C.-K., M. Dujmović, B.F.-C., N. Fiala, J.G.F., M.A.F., S.J.G., D.P.G., J. Gooty, A.G.-K., C.G., L.G., P.H.P.H., N.C.H., K.J., J.J.-N., B. Jokić, P.K., J.M.L.F., K.E.L., S.L., S.C.L., D.L., J. MacDonald, J. Matacotta, J.K.M., M. Milyavskaya, D.C.M., A. Morrow, C. Moya, H.H.N., F.N., J.Oh, T.O., R. Panczak, Y.G.P., M.F.P., K.A.Q., M. Ritchie, R.R., J.P.R., I.R., J. Samaha, S. Sankaran, D.S., L. Schnabel, A.N.S., B.A. Schuetze, A. Schütz, E.L.S., S. Shakya, M. Sirota, M.M.S., C.T.S., J.S.S., N.S., N.S.-B., C.N.H.S., E. Strømland, T. Sundelin, B.S., S. Täuber, D.T., N.v.d.V., E.J.V., J.M.V., V.V., D.W., N.W., A.L.W., B.A.N. and T.M.E. reviewed and edited the manuscript.

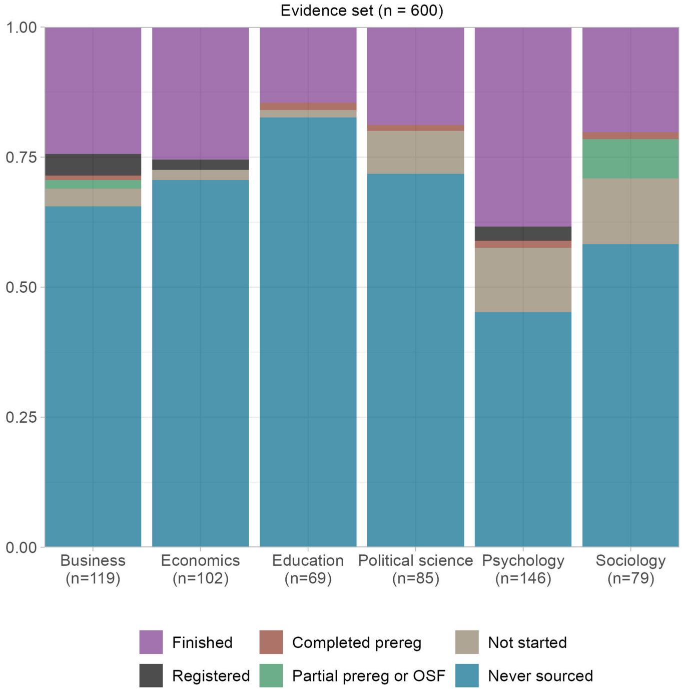
Competing interests A.H.T., M.D., N.H., K.H., O.M., T. Stankov, B.A.N. and T.M.E. are employees of the non-profit organization Center for Open Science, which has a mission to increase openness, integrity and trustworthiness of research.

Additional information

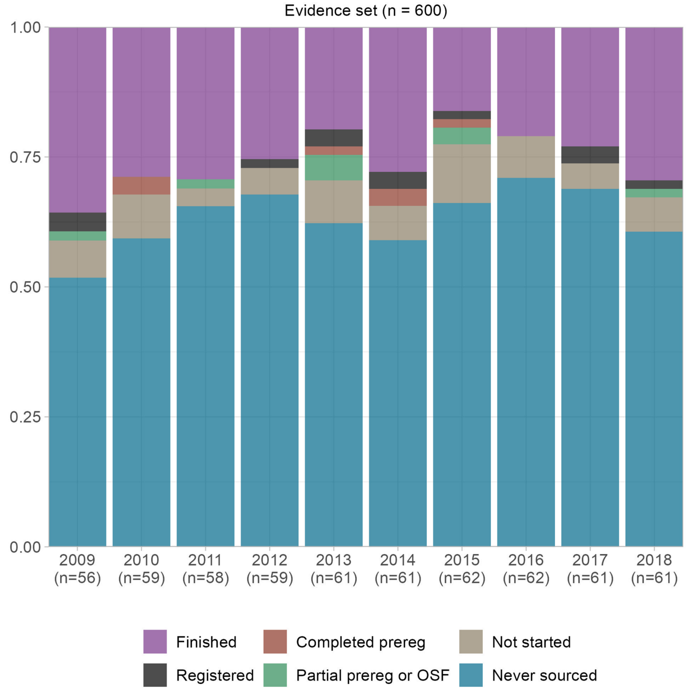
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-10078-y.

Correspondence and requests for materials should be addressed to Brian A. Nosek.

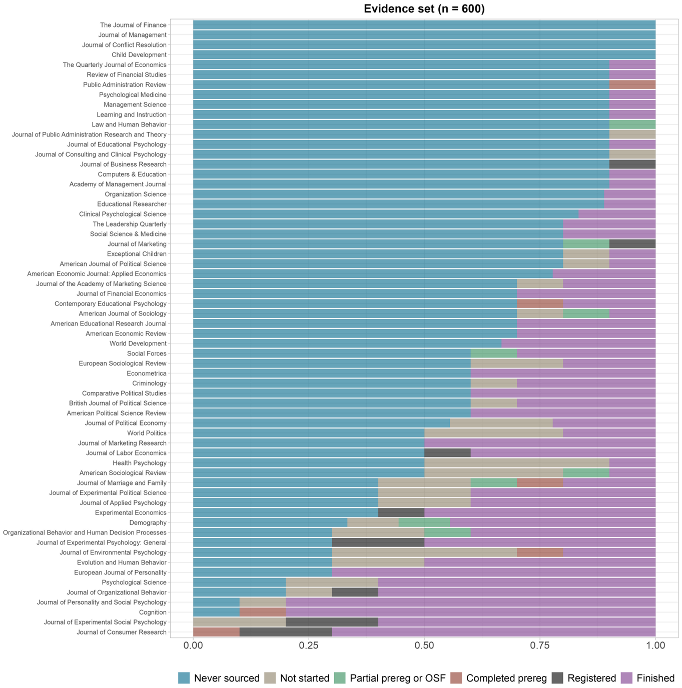
Peer review information *Nature* thanks Antica Culina, Malgorzata Lagisz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



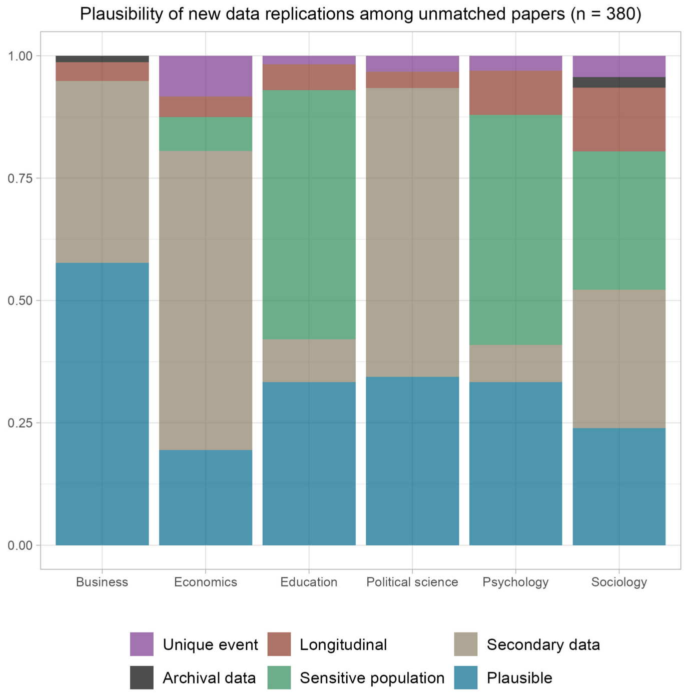
Extended Data Fig. 1 | Proportion of papers with a completed replication by discipline. Proportion of papers by discipline for which a replication attempt was finished (purple), never attempted (blue), or for which a replication team was sourced but the replication study was not started or completed (other colors). OSF = Open Science Framework. This is presented as Supplementary Fig. 7 with additional narrative context in the Supplementary Information.



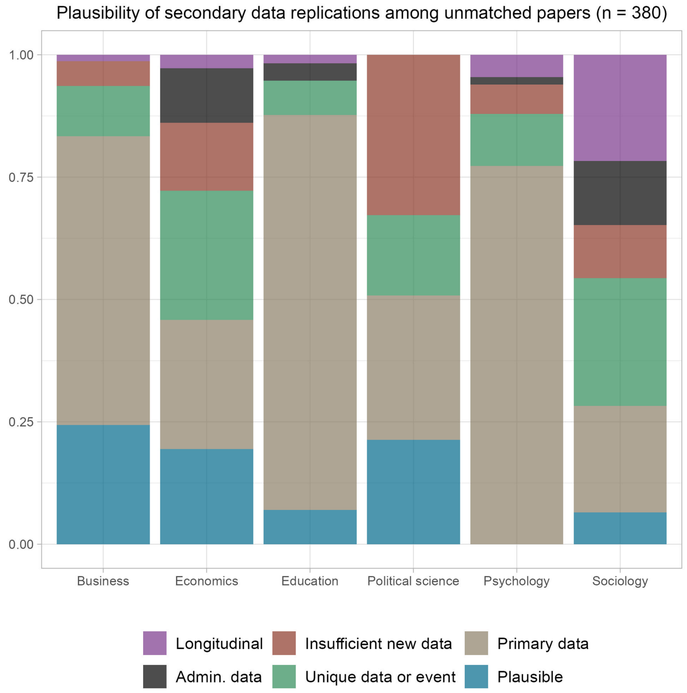
Extended Data Fig. 2 | Proportion of papers with a completed replication by year. Proportion of papers by publication year for which a replication attempt was finished (purple), never attempted (blue), or for which a replication team was sourced but the replication study was not started or completed (other colors). OSF = Open Science Framework. This is presented as Supplementary Fig. 8 with additional narrative context in the Supplementary Information.



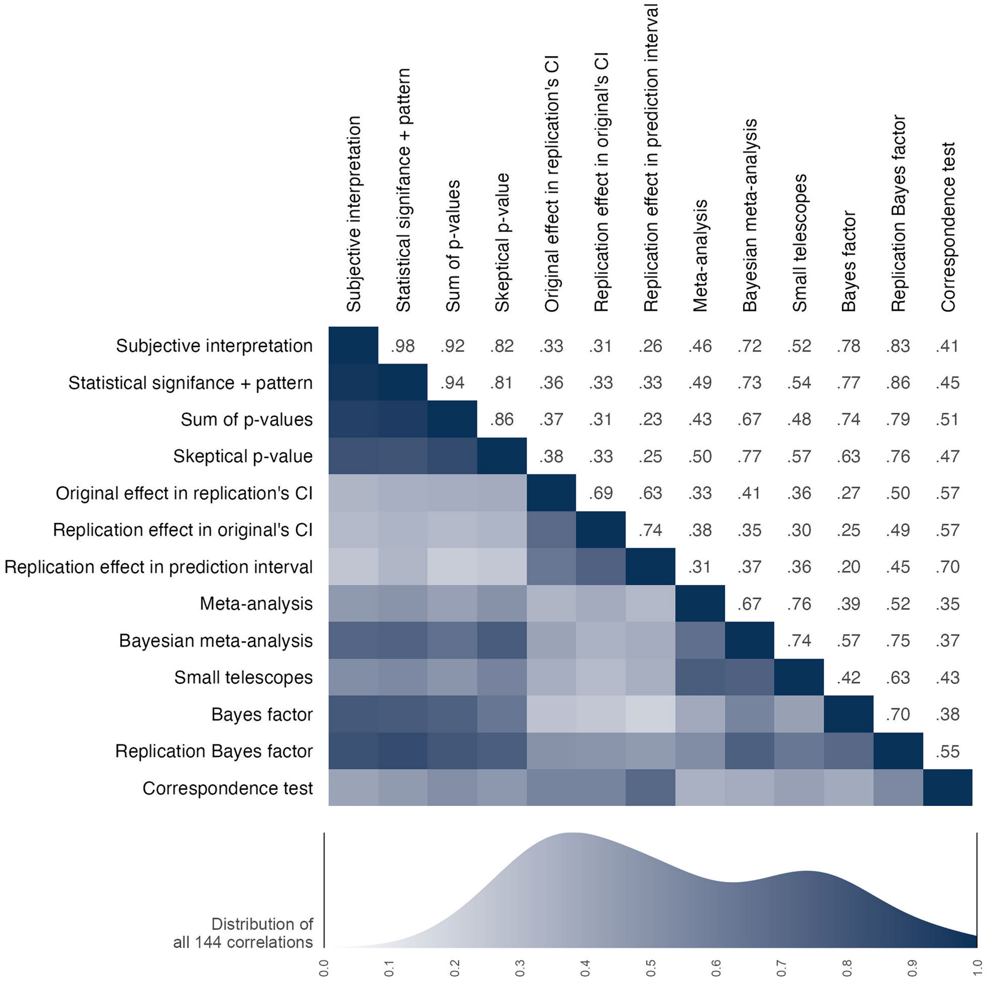
Extended Data Fig. 3 | Proportion of papers with a completed replication by journal. Proportion of papers by journal for which a replication attempt was finished (purple), never attempted (blue), or for which a replication team was sourced but the replication study was not started or completed (other colors). Sample sizes per journal ranged from 5 to 10. This is presented as Supplementary Fig. 9 with additional narrative context in the Supplementary Information.



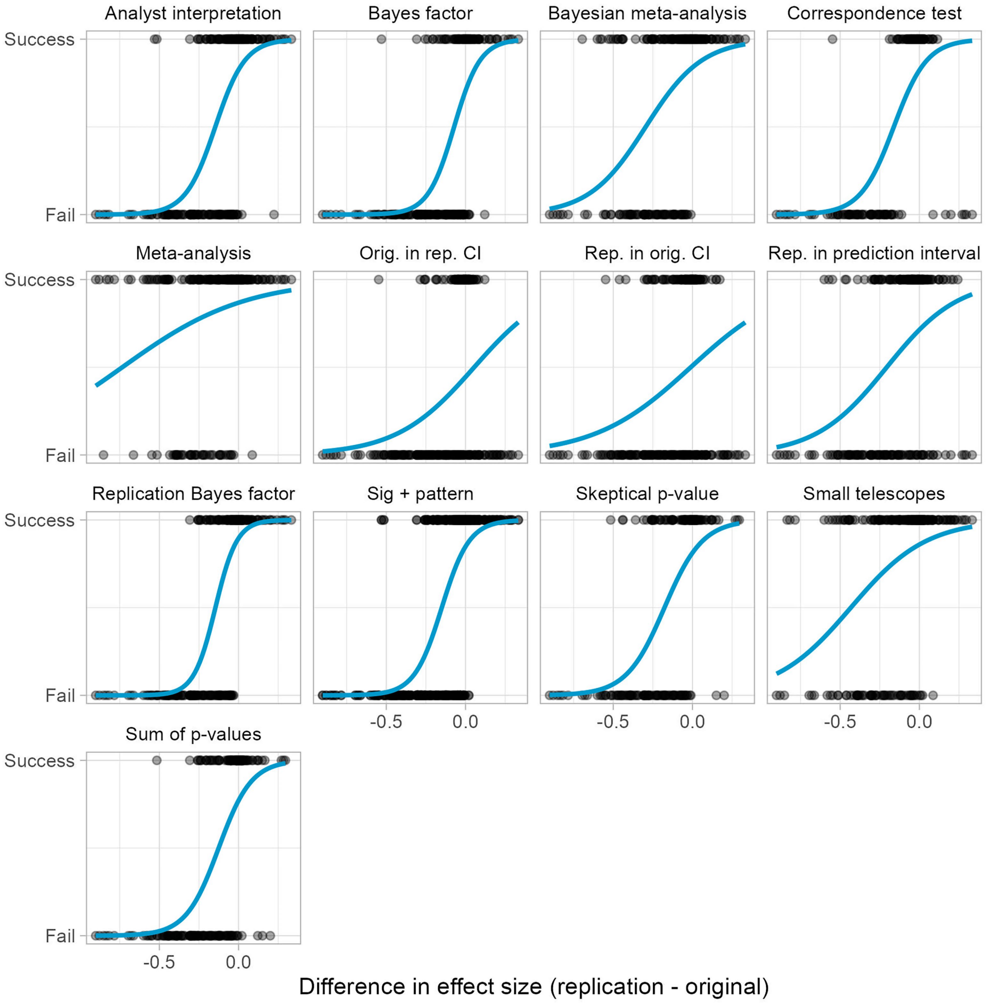
Extended Data Fig. 4 | Retrospective review of papers that were not matched to replication teams to conduct a new data replication by discipline. Y-axis indicates the proportion of available papers per discipline sample. "Plausible" means that there were no clear barriers to conducting a replication other than capacity within the project. "Secondary data" means that these papers were more appropriate for a secondary data replication. This is presented as Supplementary Fig. 10 with additional narrative context in the Supplementary Information.



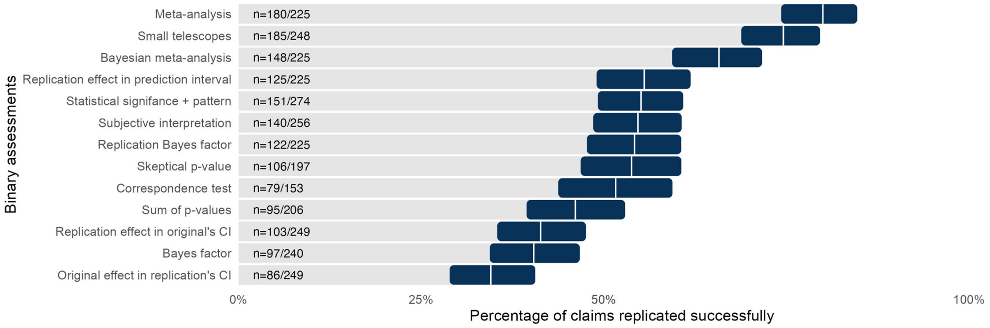
Extended Data Fig. 5 | Retrospective review of papers that were not matched to replication teams to conduct a secondary data replication by discipline. Y-axis indicates the proportion of available papers per discipline sample. "Plausible" means that there were no clear barriers to conducting a replication other than capacity within the project. "Primary data" means that these papers were more appropriate for new data replications. Admin. = Administrative. This is presented as Supplementary Fig. 11 with additional narrative context in the Supplementary Information.



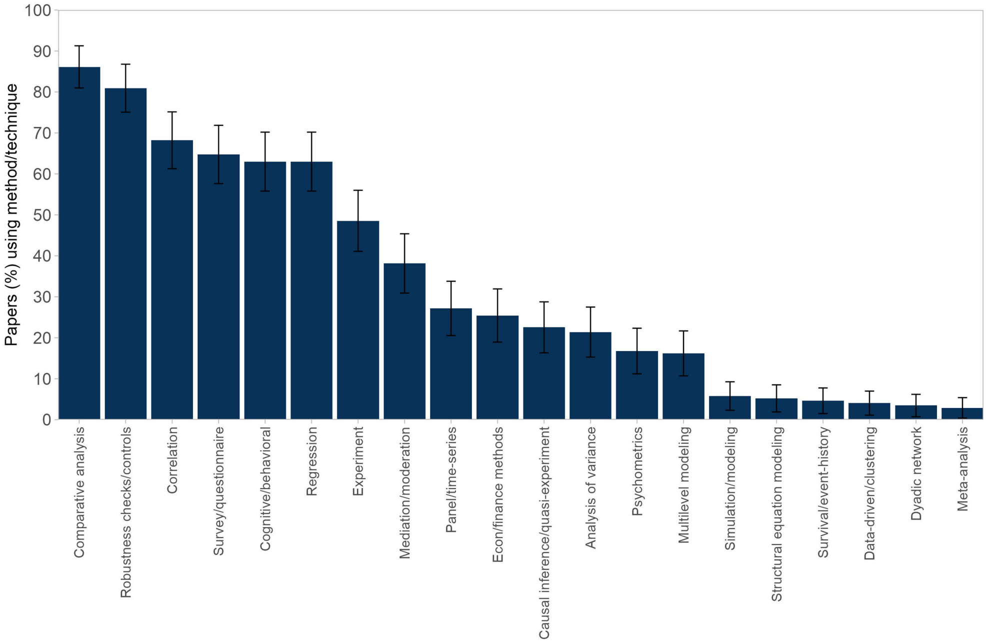
Extended Data Fig. 6 | Correlation matrix among binary assessments of replication success across claims. Correlation values are right of the diagonal, and correlation magnitude is visualized left of the diagonal with darker shading indicating stronger correlations. CI = confidence interval. This is presented as Supplementary Fig. 12 with additional narrative context in the Supplementary Information.



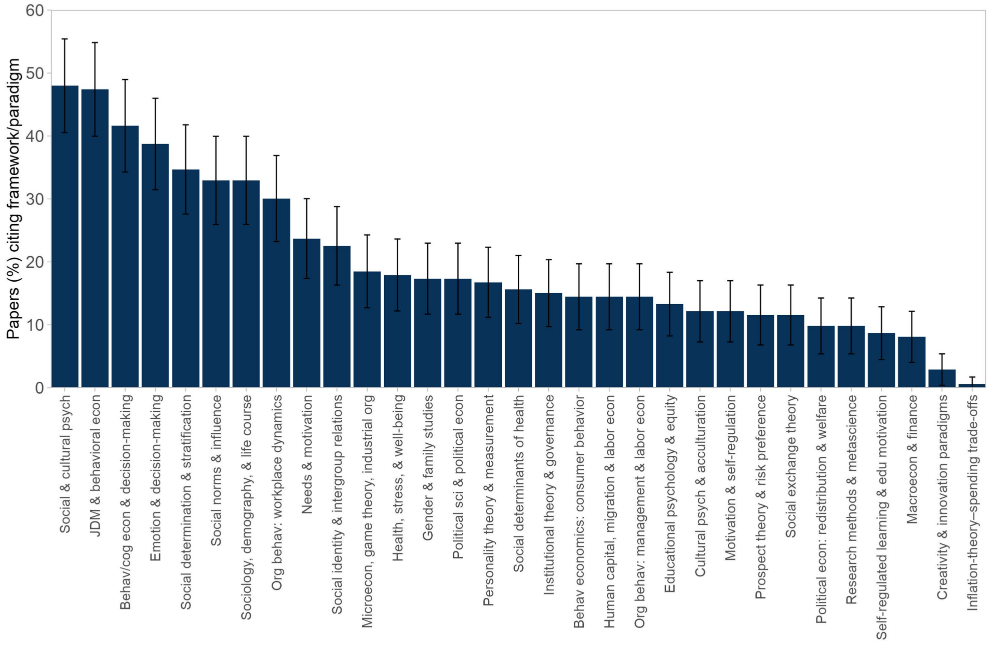
Extended Data Fig. 7 | Replication success or failure for 13 binary assessments by the effect size difference between the replication and original studies. Data points are differences in effect sizes for individual claims. Data points on top of each graph are successful replications, and data points on the bottom are failed replications, according to the graph's metric. This is presented as Supplementary Fig. 13 with additional narrative context in the Supplementary Information.



Extended Data Fig. 8 | Replication success rates across 13 binary assessments for claims. The vertical white line for each row is the estimate, and the 95% confidence interval around the estimate is represented by the dark bar. CI = confidence interval. This is presented as Supplementary Fig. 14 with additional narrative context in the Supplementary Information.



Extended Data Fig. 9 | Percentage of replicated papers that were automatically identified as using each method or technique. Two LLMs (GPT-4.1 and Kimi K2) identified the range of methods or techniques used across all abstracts (prompt: "What statistical techniques or analytic approaches are used?"). They then coded each abstract for the presence (1) or absence (0) of each—a method/technique is considered present if at least one of the models identified it as being present. Error bars = 95% confidence intervals. This is presented as Supplementary Fig. 15 with additional narrative context in the Supplementary Information.



Extended Data Fig. 10 | Percentage of replicated papers that were automatically identified as citing each theoretical framework or paradigm. Two LLMs (GPT-4.1 and Kimi K2) identified the range of frameworks or paradigms used across all abstracts (prompt: "What are the main theoretical frameworks/paradigms being cited?"). They then coded each abstract for the presence (1) or absence (0) of each—a framework/paradigm is considered present if at least one of the models identified it as being present. Error bars = 95% confidence intervals. This is presented as Supplementary Fig. 16 with additional narrative context in the Supplementary Information.

nature portfolio

Corresponding author(s): Brian Nosek

Last updated by author(s): 1/9/2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

N/A

Data analysis

All analyses of the collected replication data were performed in R (version 4.5.0). The following R packages and versions were used:

- abind [* -> 1.4-8]
- askpass [* -> 1.2.1]
- assertthat [* -> 0.2.1]
- backports [* -> 1.5.0]
- base64enc [* -> 0.1-3]
- BayesFactor [* -> 0.9.12-4.7]
- BayesRep [* -> 0.42.2]
- bayestestR [* -> 0.17.0]
- BFF [* -> 4.4.2]
- BH [* -> 1.87.0-1]
- bit [* -> 4.6.0]
- bit64 [* -> 4.6.0-1]
- boot [* -> 1.3-32]
- bridgesampling [* -> 1.1-2]
- Brobdingnag [* -> 1.2-9]
- broom [* -> 1.0.10]
- bslib [* -> 0.9.0]

```
- cachem            [* -> 1.1.0]
- callr             [* -> 3.7.6]
- checkmate         [* -> 2.3.3]
- cli               [* -> 3.6.5]
- clipr             [* -> 0.8.0]
- cluster           [* -> 2.1.8.1]
- coda              [* -> 0.19-4.1]
- codetools         [* -> 0.2-20]
- colorspace        [* -> 2.1-2]
- colourpicker      [* -> 1.3.0]
- commonmark        [* -> 2.0.0]
- contfrac          [* -> 1.1-12]
- corrplot          [* -> 0.95]
- cowplot           [* -> 1.2.0]
- cpp11             [* -> 0.5.2]
- crayon            [* -> 1.5.3]
- crosstalk         [* -> 1.2.2]
- curl              [* -> 7.0.0]
- data.table        [* -> 1.17.8]
- datawizard        [* -> 1.3.0]
- desc              [* -> 1.4.3]
- deSolve           [* -> 1.40]
- digest            [* -> 0.6.37]
- distributional    [* -> 0.5.0]
- dplyr             [* -> 1.1.4]
- DT                [* -> 0.34.0]
- effectsize        [* -> 1.0.1]
- elliptic          [* -> 1.5-0]
- evaluate          [* -> 1.0.5]
- farver            [* -> 2.1.2]
- fastmap           [* -> 1.2.0]
- fontawesome       [* -> 0.5.3]
- forcats           [* -> 1.0.1]
- foreach           [* -> 1.5.2]
- foreign           [* -> 0.8-90]
- Formula           [* -> 1.2-5]
- fs                [* -> 1.6.6]
- funkyheatmap      [* -> 0.5.2]
- gargle            [* -> 1.6.0]
- gdata             [* -> 3.0.1]
- generics          [* -> 0.1.4]
- ggExtra           [* -> 0.11.0]
- ggforce           [* -> 0.5.0]
- ggplot2           [* -> 4.0.0]
- ggridges          [* -> 0.5.7]
- ggside            [* -> 0.4.0]
- glmnet            [* -> 4.1-10]
- glue              [* -> 1.8.0]
- googledrive       [* -> 2.1.2]
- gridExtra         [* -> 2.3]
- gsl               [* -> 2.1-8]
- gtable            [* -> 0.3.6]
- gtools            [* -> 3.9.5]
- haven             [* -> 2.5.5]
- highr             [* -> 0.11]
- Hmisc             [* -> 5.2-4]
- hms               [* -> 1.1.4]
- htmlTable         [* -> 2.4.3]
- htmltools         [* -> 0.5.8.1]
- htmlwidgets       [* -> 1.6.4]
- httpuv            [* -> 1.6.16]
- httr              [* -> 1.4.7]
- hypergeo          [* -> 1.2-14]
- inline            [* -> 0.3.21]
- insight           [* -> 1.4.2]
- isoband           [* -> 0.2.7]
- iterators         [* -> 1.0.14]
- jomo              [* -> 2.7-6]
- jquerylib         [* -> 0.1.4]
- jsonlite          [* -> 2.0.0]
- knitr             [* -> 1.50]
- labeling          [* -> 0.4.3]
- lamW              [* -> 2.2.5]
- LaplacesDemon     [* -> 16.1.6]
- later             [* -> 1.4.4]
- lattice           [* -> 0.22-7]
- lazyeval          [* -> 0.2.2]
- lifecycle         [* -> 1.0.4]
- lme4              [* -> 1.1-37]
- logspline         [* -> 2.1.22]
- loo               [* -> 2.8.0]
- magrittr          [* -> 2.0.4]
- MASS              [* -> 7.3-65]
- mathjaxr          [* -> 1.8-0]
- Matrix            [* -> 1.7-4]
- MatrixModels      [* -> 0.5-4]
- matrixStats       [* -> 1.5.0]
- memoise           [* -> 2.0.1]
- metaBMA           [* -> 0.6.9]
- metadat           [* -> 1.4-0]
- metafor           [* -> 4.8-0]
- mice              [* -> 3.18.0]
- mime              [* -> 0.13]
- miniUI            [* -> 0.1.2]
- minqa             [* -> 1.2.8]
- mitml             [* -> 0.4-5]
- mnormt            [* -> 2.1.1]
- mvtnorm           [* -> 1.3-3]
- nlme              [* -> 3.1-168]
- nloptr            [* -> 2.2.1]
- nnet              [* -> 7.3-20]
- numDeriv          [* -> 2016.8-1.1]
- officer           [* -> 0.7.0]
- openssl           [* -> 2.3.4]
- ordinal           [* -> 2023.12-4.1]
- otel              [* -> 0.2.0]
- pan               [* -> 1.9]
- pandoc            [* -> 0.2.0]
- parameters        [* -> 0.28.2]
- patchwork         [* -> 1.3.2]
- pbapply           [* -> 1.7-4]
- performance       [* -> 0.15.2]
- pillar            [* -> 1.11.1]
- pkgbuild          [* -> 1.4.8]
- pkgconfig         [* -> 2.0.3]
- polyclip          [* -> 1.10-7]
- posterior         [* -> 1.6.1]
- prettyunits       [* -> 1.2.0]
- processx          [* -> 3.8.6]
- progress          [* -> 1.2.3]
- promises          [* -> 1.4.0]
- ps                [* -> 1.9.1]
- purrr             [* -> 1.1.0]
- pwr               [* -> 1.3-0]
- QuickJSR          [* -> 1.8.1]
- R6                [* -> 2.6.1]
- ragg              [* -> 1.5.0]
- rappdirs          [* -> 0.3.3]
- rbibutils         [* -> 2.3]
- RColorBrewer      [* -> 1.1-3]
- Rcpp              [* -> 1.1.0]
- RcppArmadillo     [* -> 15.0.2-2]
- RcppEigen         [* -> 0.3.4.0.2]
- RcppParallel      [* -> 5.1.11-1]
- Rdpack            [* -> 2.6.4]
- readr             [* -> 2.1.5]
- reformulas        [* -> 0.4.1]
- renv              [* -> 1.1.5]
- ReplicationSuccess   [* -> 1.3.3]
- rlang             [* -> 1.1.6]
- rmarkdown         [* -> 2.30]
- rpart             [* -> 4.1.24]
- rstan             [* -> 2.32.7]
- rstantools        [* -> 2.5.0]
- rstudioapi        [* -> 0.17.1]
- S7                [* -> 0.2.0]
- sass              [* -> 0.4.10]
- scales            [* -> 1.4.0]
- shape             [* -> 1.4.6.1]
- shiny             [* -> 1.11.1]
- shinyjs           [* -> 2.1.0]
- sourcetools       [* -> 0.1.7-1]
- StanHeaders      [* -> 2.32.10]
- stringi          [* -> 1.8.7]
- stringr          [* -> 1.5.2]
- survival         [* -> 3.8-3]
- sys              [* -> 3.4.3]
- systemfonts      [* -> 1.3.1]
- tensorA          [* -> 0.36.2.1]
- textshaping      [* -> 1.0.4]
- tibble           [* -> 3.3.0]
- tidyr            [* -> 1.3.1]
- tidyselect       [* -> 1.2.1]
- tinytex          [* -> 0.57]
- tweenr           [* -> 2.0.3]
- tzdb             [* -> 0.5.0]
- ucminf           [* -> 1.2.2]
- utf8             [* -> 1.2.6]
- uuid             [* -> 1.2-1]
- vctrs            [* -> 0.6.5]
- viridisLite      [* -> 0.4.2]
- vroom            [* -> 1.6.6]
- wCorr            [* -> 1.9.8]
- weights          [* -> 1.1.2]
- withr            [* -> 3.0.2]
- xfun             [* -> 0.53]
- xml2             [* -> 1.4.1]
- xtable           [* -> 1.8-4]
- yaml             [* -> 2.3.10]
- zip              [* -> 2.3.3]
```

All code will be publicly available when the data from this study is opened up by January 14, 2026. The stable link will be: https://osf.io/g5sny/overview

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data, materials, and code associated with this research that can be shared without restriction is available on our OSF repository. A version of record at the time of publication is registered (https://osf.io/bzfgy/; DOI: 10.17605/osf.io/bzfgy) and a living project page is also available (https://osf.io/g5sny/; DOI: 10.17605/OSF.IO/G5SNY). Also included is all available documentation for replication attempts that were not completed. This includes most of the data and code from the individual replication attempts, save for any data that is proprietary or protected that will not be made available, or for which analyst teams were uncertain or unable to confirm that they were allowed to share secondary data. It is possible that some data, materials, or code that could be shared openly is not available at the time of publication. Readers are encouraged to contact the corresponding author or the authors of the relevant subproject (Table S3) to see if more research content can be shared.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This is a quantitative study reporting data on replications of 274 claims of positive results from 164 studies.
Research sample	The sample for this study is aggregated data from replication attempts of 274 positive results from 164 original studies that were published between 2009-2018 in 62 social and behavioral science journals. There is no relevant demographic information to report as this study's unit of observation is the individual replication outcome. Original studies were made eligible for replication through stratified sampling based on the journal and the year of publication, in an attempt to collect a representative sample of articles published in these journals in this 10-year period. Original results were selected for replication by the research groups conducting the replication attempts, so the resulting sample is based on factors like the match between the original studies and the replicators' research expertise, as well as budget, timeline, and other contingencies. We conducted these replication studies to provide a sample of recent replication outcomes that cut across the core domains of the social and behavioral sciences.
Sampling strategy	We targeted stratified random selection of 30,000 papers from 62 journals across the social and behavioral sciences to be the largest sample of the published literature that could feasibly be managed with the resources of the program. We randomly selected 3,900 papers from this sample as the largest feasible sample for coding claims that could be evaluated by human and machine assessment teams. We selected 600 papers from the 3,000 as eligible for replication with recognition that costs of replication studies would limit us to conducting replications of 200 papers or fewer, so subsetting the larger sample would increase the feasibility of maintaining representativeness of the replication studies with the larger sample.
Data collection	Data collected occurred through a group of independent labs, research groups, and principal investigators who selected the original claims that they wanted to attempt to replicate. Please see the "Sourcing Replication Teams" section of the manuscript for details.
Timing	The replication teams collected data for the individual replication projects during the period of active program funding (approximately February 2019 through March 2023). We do not have the precise dates of data collection for the replication studies themselves, which were administered by the respective replication teams.
Data exclusions	19 claims from 14 papers were excluded from the final reporting. The majority of these exclusions were because of sample size concerns of the respective replication attempts. Please see the "Excluded Cases" section of the manuscript for details.
Non-participation	N/A
Randomization	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern
☒	□ Plants

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.