nature human behaviour

Article

https://doi.org/10.1038/s41562-023-01749-9

High replicability of newly discovered social-behavioural findings is achievable

Received: 25 April 2023

Accepted: 5 October 2023

Published online: 09 November 2023

Check for updates

John Protzko[1,2], Jon Krosnick[3], Leif Nelson[4], Brian A. Nosek[5,6], Jordan Axt[7], Matt Berent[8], Nicholas Buttrick[9], Matthew DeBell[3], Charles R. Ebersole[6], Sebastian Lundmark[10], Bo MacInnis[3], Michael O'Donnell[11], Hannah Perfecto[12], James E. Pustejovsky[13], Scott S. Roeder[14], Jan Walleczek[15] & Jonathan W. Schooler[1]

Failures to replicate evidence of new discoveries have forced scientists to ask whether this unreliability is due to suboptimal implementation of methods or whether presumptively optimal methods are not, in fact, optimal. This paper reports an investigation by four coordinated laboratories of the prospective replicability of 16 novel experimental findings using rigour-enhancing practices: confirmatory tests, large sample sizes, preregistration and methodological transparency. In contrast to past systematic replication efforts that reported replication rates averaging 50%, replication attempts here produced the expected effects with significance testing ($P < 0.05$) in 86% of attempts, slightly exceeding the maximum expected replicability based on observed effect sizes and sample sizes. When one lab attempted to replicate an effect discovered by another lab, the effect size in the replications was 97% that in the original study. This high replication rate justifies confidence in rigour-enhancing methods to increase the replicability of new discoveries.

Science progressively learns about the world through the discovery of replicable findings[1,2]. Efforts to systematically replicate studies across various scientific fields have reported seemingly disappointing replication rates ranging from 30% to 70%, with effect sizes (ESs) about half the size of the original findings[3–13]. This has been a cause for concern among many—but not all[14,15].

Seemingly low replicability can be the consequence of false positives or exaggerated ESs among the original studies, resulting from low statistical power, measurement problems, errors, *P*-hacking and regression to the mean due to selective reporting favouring original positive results[15,16]. Low replicability can also be the consequence of false negatives or suppressed ESs among replication studies, resulting from sampling error, low statistical power, heterogeneity of the phenomena, different analytic strategies, questionable research practices and/or lack of fidelity to the original protocols[17–22]. Finally, failures to replicate or declining ESs can be the consequence of unknown or unarticulated moderating influences and boundary conditions that differ between the original and replication studies, indicating shortcomings in the theoretical and methodological specifications[23].

Collectively, these factors comprise likely explanations for why replications are less successful and produce weaker ESs over time[24–26]. In this Article, we report the results of a prospective replication study

[1]Department of Psychological & Brain Sciences, University of California, Santa Barbara, Santa Barbara, CA, USA. [2]Department of Psychological Science, Central Connecticut State University, New Britain, CT, USA. [3]Institute for Research in the Social Sciences, Stanford University, Stanford, CA, USA. [4]Haas School of Business, University of California, Berkeley, Berkeley, CA, USA. [5]Center for Open Science, Charlottesville, VA, USA. [6]Department of Psychology, University of Virginia, Charlottesville, VA, USA. [7]Department of Psychology, McGill University, Montreal, Quebec, Canada. [8]Matt Berent Consulting, Sharon, PA, USA. [9]Department of Psychology, University of Wisconsin–Madison, Madison, WI, USA. [10]SOM Institute, University of Gothenburg, Gothenburg, Sweden. [11]McDonough School of Business, Georgetown University, Washington, DC, USA. [12]Olin School of Business, Washington University in St. Louis, St. Louis, MO, USA. [13]Educational Psychology Department, University of Wisconsin–Madison, Madison, WI, USA. [14]Darla Moore School of Business, University of South Carolina, Columbia, SC, USA. [15]Phenoscience Laboratories, Berlin, Germany. e-mail: protzko@gmail.com

Table 1 | Name, description and citation or online location of the self-confirmatory tests for the 16 discoveries included in this meta-analysis

Study name	Description of central result	Citation
Tumour	People hold others responsible for their past good behaviour caused entirely by a brain tumour but not responsible for their past bad behaviour.	https://osf.io/4n8pf/
Minimal Groups	People demonstrate less ingroup favouritism when they have been changed from one minimal group to another than when they have not changed group membership.	https://osf.io/adrbe/
Cookies	People are seen as greedier when they take three of the same kind of (free) cookie than when they take three different (free) cookies.	https://osf.io/3vz4k/?view_only=da10896b68fe4420bf6c65a3a7bd64f6
Label	When a researcher uses a label to describe people who hold a certain opinion, he or she is interpreted as disagreeing with that opinion when a negative label is used and agreeing with that opinion when a positive label is used.	https://osf.io/xq5jb/?view_only=5305a812208d48bba3e546bfe38c6c24
Self-Control	Someone who commits battery due to self-control failure is less likely to be found guilty if their failure was due to brain damage instead of genes for low self-control.	Ref. 44
Orientation	People judge same-sex interactions as more indicative of the sexuality of men than that of women.	https://osf.io/s6qdv/
Referrals	People think it is less appropriate to send product referrals when they are the sender of the referral than when they are merely judging the behaviour of someone else who sends the referral.	https://osf.io/v3thd/?view_only=1c32808a01ee4c8c816480825ad5bebf
Ads	Watching a short ad within a soap-opera episode increases one's likelihood to recommend and promote the company in the ad.	https://osf.io/ngz5k/?view_only=8cf18a2babc1499e98ef57dbb9926a80
FSD	Forcing people to answer questions quickly makes them give more socially desirable answers.	Ref. 45
Prediction	People make more complicated sets of predictions when asked to do so without having the opportunity to explore data.	https://osf.io/e2sf8/
Fairness	People evaluate the fairness of punishments that can be expressed in multiple currencies (for example, time and money) on the basis of whichever currency is initially more salient, but they update their fairness judgements when the translation to the other currency is highlighted.	Ref. 46
Ostracism	People who are ostracized by computers in an online ball-tossing game become less trusting towards people in general.	https://osf.io/58vz9/?view_only=312b6136155849a79f3416933a05789b
Misattribution	People misattribute the feeling of a-ha! they feel while solving an anagram to the truth of the statement the anagram is embedded in.	Ref. 47
Redemption	People who look different after committing a transgression seem more remorseful, trustworthy and deserving of a second chance than people who look the same as they did at the time of their transgression.	http://osf.io/6h5s2/
Worse	People evaluate Team B more favourably when they are told that Team A is more qualified than Team B than when they are told that Team B is less qualified than Team A.	Ref. 48
Misreporting	People will report not engaging in an activity if they learn that reporting they have engaged in the activity will result in several additional questions.	https://osf.io/3ud4s/?view_only=420da5b8113b42be8710cd7c4b4af14a

examining whether low replicability and declining effects are inevitable when using proposed rigour-enhancing practices.

Four laboratories conducting discovery-oriented social-behavioural research participated in a prospective replication study (Supplementary Information section 1). Over five years, the labs conducted their typical research, examining topics covering psychology, marketing, advertising, political science, communication, and judgement and decision-making (Table 1). Each lab engaged in pilot testing of new effects based on their laboratory's business-as-usual practices. These practices could involve collecting data with different sample providers and with any sample size the lab saw fit. All pilots were required to have their materials, procedure, hypotheses, analysis plan and exclusions preregistered prior to data collection. Ostensible discoveries were then nominated for confirmation and replication by the discovering labs. The main criterion for moving from piloting and exploration into the confirmation and replication protocol was that the lab believed they had discovered a new effect that was statistically distinguishable from zero during the piloting phase. Each of the four labs submitted four new candidate discoveries for a self-confirmatory test and four replications, for a total of 16 confirmatory tests and 64 replications. In the self-confirmatory test, the discovering lab conducted a preregistered study with a large sample ($N \geq 1,500$) and shared a report of the methodology. Regardless of the outcome of the self-confirmatory test, in the replication phase, all four labs conducted independent preregistered replications using the written methodology and any specialized study materials shared by the discovering lab (for example, videos constructed for delivering interventions). Ordinarily, we would promote strong communication between labs to maximize sharing of tacit knowledge about the methodology, but in this case, to maintain the independence of each replication, we opted to discourage communication with the discovering lab outside of the documented protocols except for critical methodology clarifications (Supplementary Information section 2). The replicating labs used equally large sample sizes (all $N \geq 1,500$), and each lab used a different sample provider.

Preregistration, reporting all outcomes, large sample sizes, transparent archiving, sharing of materials and commitment to high-fidelity replication procedures should minimize irreplicability or declining ESs stemming from questionable research practices, selective reporting, low-powered research or poorly implemented replication procedures. Such optimizing might promote higher replicability than previously reported in the literature. If—despite these rigour-enhancing practices—low replicability rates or declining effects are observed, such rates or declines could be intrinsic to social-behavioural scientific investigation[24,25,27–30].

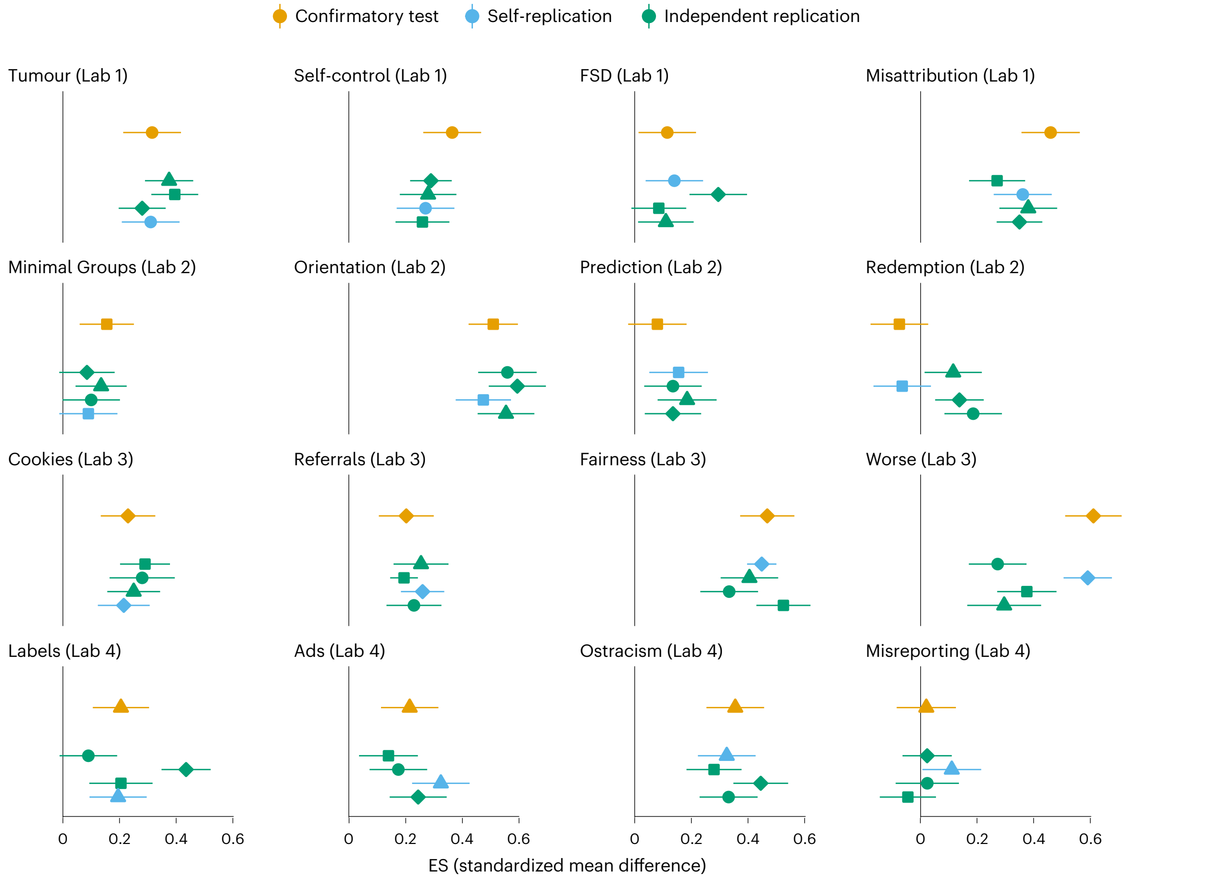

Fig. 1 | Effect size estimates and 95% CI from 16 new discoveries in the social-behavioral sciences, with four replications each. ESs (shapes) and 95% CIs from 16 new discoveries (yellow) in the social-behavioural sciences with four replications each. Each lab is designated by a unique shape for the observed ES; blue marks correspond to self-replications, green marks to independent replications.

Each of the 16 ostensible discoveries were obtained through pilot and exploratory research conducted independently in each laboratory. Not every pilot study the labs conducted was put forward for confirmation and replication. Like all exploratory research, labs sometimes found errors, did not find signals of potential effects or just lost interest in pursuing it further. The labs introduced 4 provisional discoveries each, resulting in 16 self-confirmatory tests and 64 replications (3 independent and 1 self-replication for each), testing replicability and decline. All confirmatory tests, replications and analyses were preregistered both in the individual studies (Supplementary Information section 3 and Supplementary Table 2) and for this meta-project (https://osf.io/6t9vm).

Results

Replicability rate

Self-confirmatory tests. Of the 16 discoveries put forward for replication, 81% (13/16) produced statistically significant results during the self-confirmation phase Effect Size ($\bar{d}$ = 0.27; $t(15) = 5.61$; $P < 0.001$; 95% confidence interval (CI), 0.17 to 0.37; estimated between-study heterogeneity s.d., 0.18). The average ES of the self-confirmatory tests was smaller than the estimated average ES of the published psychological literature ($\bar{d}$ = 0.43)[31], even when considering only the 13 statistically significant findings (simple average $\bar{d}$ = 0.32). No lab produced self-confirmatory tests with larger average ESs than the other labs (robust approximate Hotelling's T^2(6.01), 0.60; $P = 0.64$).

Replications. One way of assessing replicability is to examine whether each replication rejects the null hypothesis at $P < 0.05$ in the expected direction[6]. Including all 16 self-confirmatory tests, 55/64 (86%) replications were successful. The average ES of the replications was $\bar{d}$ = 0.26 ($t(15) = 7.61$; $P < 0.001$; 95% CI, 0.19 to 0.33), similar to the ES observed in the self-confirmatory tests ($\bar{d}$ = 0.27; $t(15) = 5.61$; $P < 0.001$; 95% CI, 0.17 to 0.37) yet larger than replication ESs observed in prior attempts to systematically replicate the established social-behavioural literature ($\bar{d}$ = 0.16)[32].

Focusing only on the 13 statistically significant self-confirmatory tests, 47/52 (90%) of the subsequent replications yielded statistically significant effects in the hypothesized direction. The average ES of the replications was $\bar{d}$ = 0.32, similar to the ES observed in the 13 self-confirmatory tests ($\bar{d}$ = 0.32).

For the three statistically non-significant self-confirmatory tests, 8/12 (67%) of the subsequent replications produced statistically significant effects in the hypothesized direction: 1/4 for one ($d = 0.03$), 3/4 for the second ($d = 0.09$) and 4/4 for the third ($d = 0.15$). These replication ESs were slightly larger than the ES observed in the original self-confirmatory tests ($\bar{d}$ = 0.01; Fig. 1).

An alternative index of replicability examines the consistency of ESs generated by the initial self-confirmatory test and its subsequent replications (within-study heterogeneity; $\hat{\tau}_{\text{within}}$). On the basis of a multilevel meta-analysis, little variation in ESs was observed beyond what would be expected by sampling variation alone ($\hat{\tau}_{\text{within}} = 0.06$; profile likelihood 95% CI, 0.04 to 0.08; $P < 0.001$). This indicates that the self-replications and independent replications would not have perfectly replicated the ESs of the self-confirmatory tests, even if all samples were large enough to make sampling error negligible. The degree of variation was, however, smaller than the variation in ESs across the 16 discoveries ($\hat{\tau}_{\text{between}} = 0.14$; profile likelihood 95% CI, 0.1 to 0.2; $P < 0.001$). This suggests that the variation due to sampling and procedural differences across labs was much smaller than the variation due to the phenomena being investigated[7,8].

In an exploratory model, differences between self-confirmatory tests and self-replication ESs were found to be fully attributable to sampling error, and average ESs in independent replications strongly correlated with ESs from self-confirmatory tests and self-replications ($r = 0.83$; profile likelihood 95% CI, 0.52 to 0.95; $P < 0.001$). There was a small amount of heterogeneity across the independent replications ($\hat{\tau}_{\text{within}} = 0.05$; 95% CI, 0.03 to 0.08; $P < 0.001$). This suggests that conducting replications in new samples and with independently implemented methods increased variability in the observed ESs but did not systematically reduce the observed ESs.

On the basis of a power analysis of the 13 self-confirmatory tests with statistically significant results, the average replication power was 0.96 with a median approaching 1 and an average power in replication studies of specific discoveries ranging from 0.62 to approaching 1. The observed replication rate of 90% is slightly smaller than expected on the basis of these power estimates. Considering all self-confirmatory tests (including statistically non-significant ones), the average power across attempted replications was 0.80 with a median of 0.99 and a range of 0 to approaching 1. The observed replication rate of 86% in the replications for all 16 discoveries was somewhat larger than expected on the basis of these power estimates. Overall, replication rates were consistent with power estimates calculated using the ESs observed in self-confirmatory tests.

Declines in ESs across replications

Within-study heterogeneity across replications was estimated to be s.d. = 0.06, suggesting little heterogeneity overall, despite 75% of the replications being conducted independently using different sample providers. There was modest evidence that one lab produced slightly smaller ESs in replications than did one other lab, controlling for the average size of effects in the initial self-confirmatory tests from each lab (robust approximate Hotelling's $T^2(12.31)$, 3.51; $P = 0.05$).

When comparing the self-confirmatory tests to their self-replication attempts, no significant evidence of declining effects appeared. No originating lab's self-confirmatory test ES was significantly different from its self-replication attempt (Fig. 2). On average, self-replications had the same ES as the self-confirmatory tests ($\underline{d}_{\text{difference}} = -0.00$; $t(15) = -0.15$; $P = 0.88$; 95% CI, −0.03 to 0.03; between-study heterogeneity $\hat{\tau} = 0.00$; profile likelihood 95% CI, 0.00 to 0.06; $P = 1.00$). In three cases, however, there was a statistically significant difference between the ES in the self-confirmatory test and the average ES across independent replications. As both positive and negative discrepancies occurred, independent replication ESs were the same size as those of self-confirmatory tests, on average ($\underline{d}_{\text{difference}} = -0.01$; $t(15) = -0.21$; $P = 0.84$; 95% CI, −0.06 to 0.05). The discrepancies were heterogeneous across studies, with an estimated between-study s.d. of $\hat{\tau} = 0.09$ (profile likelihood 95% CI, 0.06 to 0.16; $P < 0.001$). On the basis of this degree of heterogeneity, independent replications of new discoveries would be predicted to differ from the ES of the self-confirmatory test by as much as s.d. = 0.2 (95% prediction interval, −0.20 to 0.21).

The results of half of the self-confirmatory tests and replications were blinded from data analysis until all replications were completed, to test whether awareness of outcomes influenced replication success and/or the potential declining of ESs over replications. Whether the studies' results were blinded did not moderate the results (Unstandardized Regression Coefficient, $b = -0.01$; $t(14) = -0.26$; $P = 0.80$; 95% CI, −0.07 to 0.06 for differences between self-confirmatory tests and self-replications; $b = 0.07$; $t(14) = 1.34$; $P = 0.20$; 95% CI, −0.04 to 0.19 for differences between self-confirmatory tests and independent replications). Finally, when we tested ESs sequentially over time, no evidence for a decline in ESs from the self-confirmatory test through the final replication was observed ($b = -0.002$; $t(73) = -0.38$; $P = 0.71$; 95% CI, −0.02 to 0.01; Fig. 3). These results did not significantly change when we removed the fixed effect for each lab (for all additional and robustness checks, see Supplementary Information section 4). In addition, the null was not rejected when we compared the slopes of change in ESs among 'blind' versus 'not blind' studies ($b = 0.02$; $t(73) = 1.75$; $P = 0.10$; 95% CI, −0.01 to 0.05).

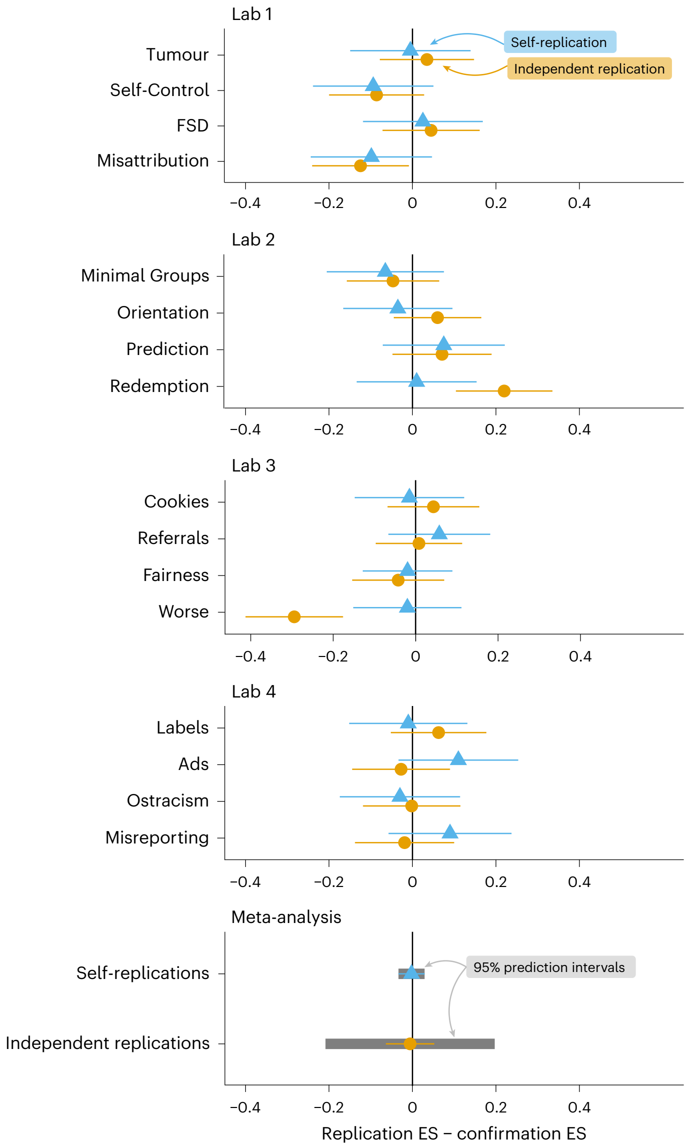

Fig. 2 | Difference in effect size estimates between self- and independent replications for 16 new discoveries in the social-behavioural sciences, compared with initial confirmation study effect size estimates. Difference in ESs (shapes) between self-replications and independent replications for the 16 discoveries, compared with the self-confirmatory test ES (0.0 on the *x* axis). The independent replication estimates are fixed-effects meta-analytic estimates of the three independent studies. The meta-analytic estimates with prediction intervals in the bottom panel combine across all 16 discoveries.

An uninteresting reason for high replicability would be if the discoveries, although novel, are obviously true. Trivial findings might be particularly easy to replicate. To assess this, we conducted two additional studies (Supplementary Information section 5; the protocol (no. 156-19-0689) was deemed exempt by the Office of Research on Human Subjects (Institutional Review Board) at the University of California, Santa Barbara). In the first study, 72 researchers reviewed a synopsis of most of the research designs and predicted the direction of each finding. On average, raters correctly predicted the direction and significance of the self-confirmatory tests 42% of the time,

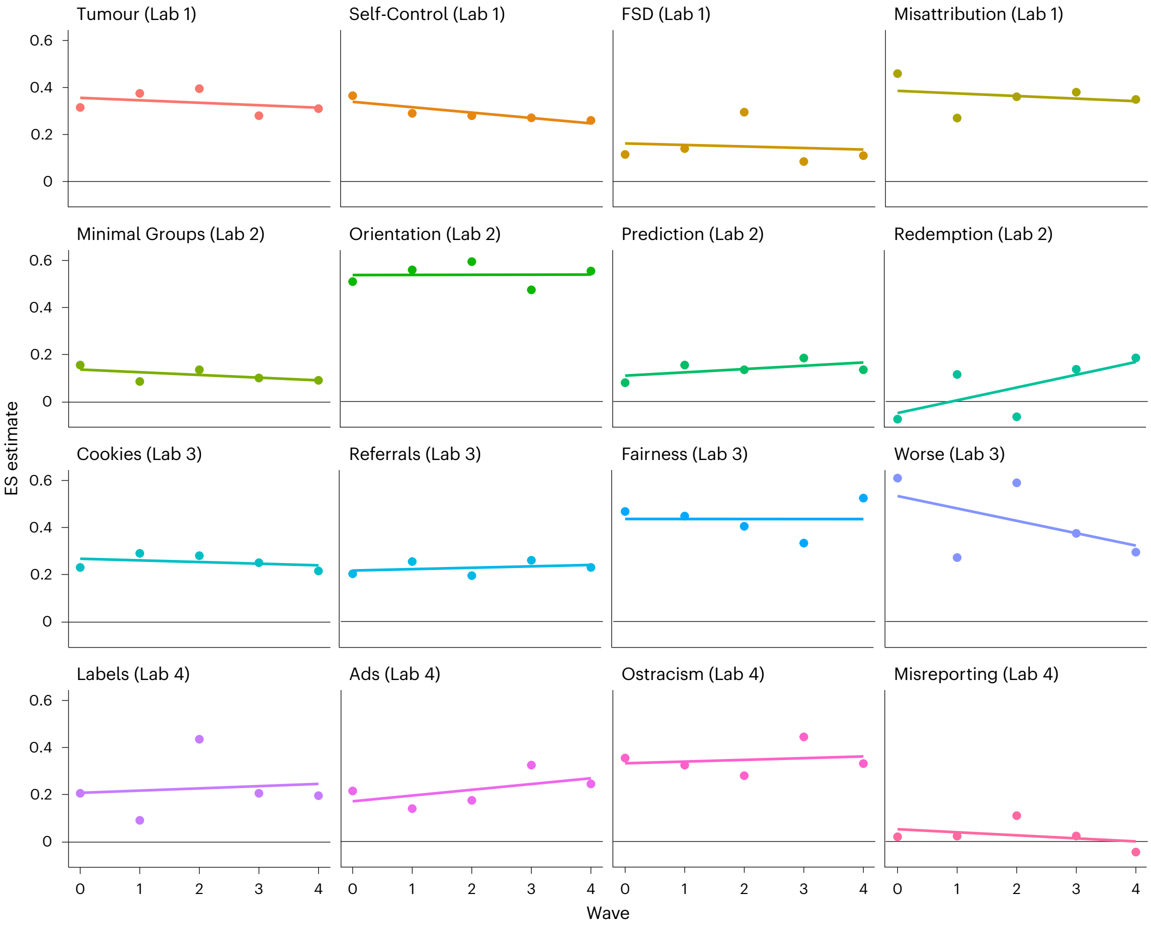

Fig. 3 | Changes in effect size across replications, with initial confirmation study at the intercept. Slopes of ES changes across replications, with the self-confirmatory test as the intercept.

incorrectly predicted null results 38% of the time and incorrectly predicted the direction of the findings 20% of the time. In the second study, 1,180 laypeople reviewed synopses of the research designs from this study, which showed high replicability, and from a prior study of published findings from the same fields with similar methodologies that showed low replicability[6,8,9]. The synopses were generated by independent researchers with experience in designs using synopses. On multiple preregistered criteria, the participants were no better at predicting the outcomes of the highly replicable discoveries presented here ($\text{mean}_{\text{present studies}} = 41\%$ correct prediction) than at predicting the other less replicable findings from the prior investigation ($\text{mean}_{\text{comparison studies}} = 42\%$; $\hat{\Delta} = -1.65$; 90% CI, −3.4 to 0.1; $t(1,172) = -6.32$; $P < 0.001$ for a preregistered equivalence test of the null hypothesis that the difference between present studies and comparison studies would exceed 5 percentage points, H_0: $\Delta \geq 5.00$). Notably, the average accuracy rate of researchers in the first study was nearly identical to the average accuracy among laypeople in the second. Additionally, the accuracy of predictions for specific findings was significantly associated with the absolute magnitude of the average ESs from independent replications ($b = 2.79$; $z = 2.95$; $P = 0.003$; 95% CI, 0.94 to 4.66 for the findings in the present study; $b = 0.66$; $z = 3.05$; $P = 0.002$; 95% CI, 0.24 to 1.09 for the comparison findings); absolute ES explained 35% of the variance in predictability rates. These findings indicate that the sample of discoveries used here were not of a prima facie different type of content that would yield high replication rates. Nor were the content or hypotheses more obvious or predictable than similar findings with low replication rates.

Discussion

Rather than beginning with published findings and attempting to replicate them in a retrospective replication investigation, we implemented a prospective investigation[33]. By discovering new experimental effects, we were able to apply a meta-scientific[34] lens to the entire process from discovery to a confirmatory test, and through a sequence of replication attempts. By subjecting ostensible discoveries to large-sample, preregistered confirmatory tests (13/16; 81% supported at $P < 0.05$; $d = 0.27$), we ensured that the primary findings for replication were free from P-hacking and questionable research practices, unlikely to be artefacts of low statistical power, and fully documented. This was true not only of the original discoveries of new effects but also in their replications.

These 16 discoveries, four from each of four independent labs, were then subjected to independent sequential replication (55/64; 86% supported at $P < 0.05$; $d = 0.26$). Considering replications of only the 13 statistically significant self-confirmatory tests, the observed replicability rate was 90%. Considering the power to detect ESs from all 16 self-confirmatory tests, the replication rate could not be any higher. The replication ESs were the same size on average as those of the self-confirmatory tests when conducted by the discovering labs and 97% the ESs found by independent labs. Prior replication efforts in the social-behavioural sciences report replication success rates of about 50% on average, producing ESs less than half of the originally reported ESs[3–13,32]. The present findings establish a benchmark showing that high replicability in the social-behavioural sciences is achievable in both statistical inference and ES estimation, when the original studies and replications are conducted using ostensibly best practices and replications are conducted with complete fidelity.

We investigated whether low replicability and declining ESs should be expected from the social-behavioural sciences because of the complexity of the phenomena, hidden moderators[18] and other factors that might be intrinsic to the phenomena being studied or to the replication process[24,25]. Instead, we found a high replicability rate. The present results are reassuring about the effectiveness of what we think of as best practices in scientific investigations. When novel findings were transparently subjected to preregistered, large-sample confirmatory tests—and when replications involved similar materials and were implemented with a commitment to faithfulness to testing the same hypothesis with fidelity to the original procedure—the observed rate of replication was high. Furthermore, we saw no statistically significant evidence of declining ESs over replications, either when holding materials, procedures and sample source constant (except for sampling error) or when materials, procedures and sample sources varied but were faithful to the original studies.

We further showed, with one survey of scientists and one survey of the lay public, that our studies were not a priori more predictable than similar other studies with a known replication rate. Our high replication rate thus cannot be attributed to our studies being different in especially replicable topics, designs or hypotheses. It would also be possible to imagine that we observed higher replicability than other investigations because of the qualities of the researchers involved in this project, such as being better at imagining and discovering new, replicable phenomena. Although we could be motivated to believe this possibility, the principal investigators in this project all have direct experience with their own published findings failing to replicate. Also, in this and other research, the participating labs have established practices of making risky predictions, most of which fail to materialize into reliable phenomena. If there is an investigator influence on the observed findings, we believe that it is aligned with our interpretation of the present evidence as being due to the adoption of rigour-enhancing practices as lab norms rather than individual exceptionality[35].

It is likely that we observed high replicability because of the rigour-enhancing methodological standards adopted in both the original research leading to discovery and the replication studies. First, rather than using exploratory discoveries as the basis for claiming a finding, all discoveries were subjected to preregistered self-confirmatory tests. This eliminated inflation of false positives and ESs by pre-commitment to research designs and analysis plans[36]. Second, once a discovery was submitted for a self-confirmatory test, we committed to reporting the outcomes. This eliminated publication bias, which is particularly pernicious when selective reporting of study findings systematically ignores null results[37–39]. Third, all self-confirmatory tests and replications were conducted with large sample sizes ($N \gtrsim 1{,}500$), resulting in relatively precise estimates. Fourth, each lab was part of the process of both discovering and replicating findings. This may have motivated teams to be especially careful in both characterizing their methods and carrying out their replications. Fifth, if there were essential specialized materials for the experimental design, the discovering lab made them available as supplementary materials. Sharing original materials should increase understanding of and adherence to critical features of original experimental methodologies. We expect that all these features contributed to improving replicability to varying degrees. Future investigations could manipulate these features to learn more about their causal contributions to replicability.

Even when using rigour-enhancing processes, independent replications conducted by other laboratories produced effect sizes that differed from self-confirmatory effect sizes by larger margins than would be expected due to sampling error alone, producing both larger and smaller ESs with a heterogeneity on the order of s.d. = 0.2. The added heterogeneity may be due to ambiguities or imprecise descriptions in the materials provided to replicating labs, or due to sample heterogeneities introduced by using varying sample providers. This finding highlights the value of multi-lab replication processes even when the original studies follow rigorous practices.

An important question is the constraints on the generalizability of these findings to other research. Our 16 novel findings in social-behavioural sciences each involved two between-subject conditions that could be administered online. All samples for the self-confirmatory tests and replications were drawn from online panels of American adults. These 16 findings do not characterize a representative sample of any methodology or discipline, although they do represent common methodologies, samples and research questions from the social-behavioural sciences. Due to the limited number of participating labs, lab-level variation in the replicability of findings was incalculable; to the extent that labs vary in how they select potential replication targets, the replication rates observed in the present study may not generalize to a broader population of research groups, although—as discussed above—the discovered effects did not differ from similar findings when asking laypeople and a group of researchers to predict them.

The most obvious areas for further testing of the replicability ceiling include more complex experimental or observational designs, in-person behavioural designs that have intricate staging or unusual protocols, sampling strategies that are more inclusive of the world's population, more heterogeneous conditions including variations in procedure and time between investigations, and research domains beyond the social-behavioural sciences. It is possible that our computer-administered protocols are inherently more replicable, though prior replication efforts of similar methodologies suggest that this is not the case[8,9,12], and our follow-up investigation found that the present findings were no more predictable a priori than other findings with similar methodologies that frequently failed to replicate (Supplementary Information section 5). Systematic investigation will be fruitful for understanding the boundary conditions for achieving high replicability.

Low replicability or declining ESs in social-behavioural research are not inevitable. We did not observe declining effects due to idiosyncrasies of different laboratory practices or different sampling conditions. Whereas prior research demonstrates that such declines can occur, the present research demonstrates that they do not necessarily occur. The encouraging evidence here should empower scientists with confidence that what we believe to be rigour-enhancing practices may indeed efficiently generate reliable insights. With the adoption of rigorous research practices such as confirmatory testing, large samples, preregistration, strong documentation and fidelity of replication, high replicability is achievable—perhaps even likely.

Methods

All research complies with all relevant ethical regulations; all studies were approved by the local lab's Institutional Review Board. The four participating labs conducted pilot and exploratory research in the social sciences pursuing their own typical practices and research interests independently of the other labs. The labs were encouraged to investigate any aspect of social-behavioural science, with the requirements that the discoveries submitted for self-confirmatory testing and replication be novel and non-obvious and not involve deception.

The labs submitted promising discoveries for self-confirmatory tests and replication if they met the following inclusion criteria: a two-group between-subjects manipulation with one focal dependent variable, with methods administered via computer online to adults within a single 20-minute study session.

Prior to conducting the self-confirmatory test, the discovering labs preregistered the study design and analysis plan, including materials, protocol, plans for data cleaning and exclusion, and specification of the analysis model. Once a self-confirmatory test was preregistered, the lab wrote a methods section to share with the other labs. These methods sections had to include everything that the discovering lab believed would be required for an independent lab to conduct an effective replication. This was done to capture the naturalistic conditions when a researcher reads a methods section and conducts a replication based on it.

Following preregistration, no changes could be made to the methods or procedures, and all labs were committed to replicating the protocol regardless of the outcome of the self-confirmatory test. The discovering lab conducted its self-confirmatory test with about 1,500 participants, and then the project coordinator initiated the replication process with the other labs. The labs were assigned the order to conduct replications in a Latin square design to equate lab-specific effects across the order of replications (Supplementary Information section 6).

Sharing study descriptions

After a lab identified an ostensible discovery for a self-confirmatory test, they distributed a description of the methodological details that they believed would be required for an independent lab to run a replication. When the replicating labs considered the instructions to be ambiguous on a meaningful part of the design (71% of studies), the replicating labs sought clarifications about methodology from the discovering lab. Usually these were trivial clarifications or confirmations, but not always (Supplementary Information section 2).

Replications were done sequentially following the same protocol as the self-confirmatory tests, including preregistration. Variation from the 1,500 participants per study was due to idiosyncrasies in how the panels and labs managed participant flow and the application of preregistered exclusion criteria. In most cases, the panels allowed more participants to complete the questionnaire.

The discovering labs could specify required exclusion criteria, such as attention checks. The replicating labs could also choose to preregister and implement exclusions for attention checks following their own laboratory's best practices. This was done to capture the natural way researchers conduct replications using their own view of best practices. To maintain the ecological validity of labs conducting research in their own style, and to maximize the independence of each replication, all sharing of materials was managed by a project coordinator to prevent unintended communication of designs or results.

Main studies

Sixteen new discoveries of social-behavioural phenomena were submitted to self-confirmatory testing and replication, four from each of the participating laboratories. Table 1 catalogues the new discoveries with a brief name, a one-sentence summary of the finding and a citation to the research. Supplementary Table 3 provides links to comprehensive information for each self-confirmatory test and replication, including the preregistration with the design and analysis plan, research materials, data, analysis code, analysis output and written reports of the methods and results.

Participants. The population of interest for the self-confirmatory tests and replications was adults living in the United States who could read and write in English. The participants were members of panels that had been recruited through non-probability sampling methods to complete online questionnaires in return for small amounts of money or redeemable 'points'[40,41]. Labs contracted with different sample providers to provide participants (Stanford University: Toluna, SSI and Dynata; University of California, Santa Barbara: CriticalMix; University of California, Berkeley: Luth; University of Virginia: SoapBox Sample and Lightspeed GMI). We used different sample providers to minimize potential overlap in sampling, although we cannot be sure that some participants are not part of multiple panels and also repeated our studies as part of different panels. These samples were taken from the providers' online, opt-in, non-probability panels. The sample providers were instructed to provide American adults drawn in a stratified way with unequal probabilities of selection from the panels so that the people who completed each survey would resemble the nation's adult population (according to the most recently available Current Population Survey, conducted by the US Census Bureau) in terms of gender, age, education, ethnicity (Hispanic versus not), race (allowing each respondent to select more than one race), region and income. This method produced samples designed to look similar to probability samples on the matched characteristics, but the samples may still have differed in unknown ways on unmatched characteristics. The sample providers may have varied in their success at achieving representativeness. A potential lack of adherence to that sampling plan was non-consequential for the conducted studies. For none of the discoveries were the findings presumed to be limited to a subsample of adults, although there may have been a priori or post facto hypothesizing about moderation by demographic variables. For the pilot and exploratory studies, the labs used whatever samples they wished (for example, panel, MTurk or participants visiting the laboratory).

Blinding and sample-splitting manipulations. Two planned manipulations of secondary interest were included to explore potential reasons for variation in the replicability rate or its decline over time. One involved randomly assigning participant recruitment for each data collection of 1,500 participants into a first and second wave of 750 to investigate declines in ES across a single data collection. We assign less confidence to this manipulation, however, as not all panels may have consistently followed our strict protocols for this random assignment (see Supplementary Information section 7 for all additional procedures that the labs and sample providers were instructed to follow). The second manipulation randomly assigned 8 of the 16 new discoveries (2 from each team) to blind the results of the primary outcome variable from the self-confirmatory tests and replications for all team members until all replications for that finding had been completed. For the other 8 discoveries, the data were analysed and reported to the other teams as the results became available. This was to determine whether explicitly blinding research findings would moderate replicability rates and/or declining ESs across replications[24,25].

Confirmatory analysis

Meta-analysis. In all analyses, meta-analytic models estimated with restricted maximum likelihood were used, as implemented in the metafor package (version 4.2-0) for R version 4.2.2 (ref. 42,43). For single-level models, Knapp–Hartung corrections for standard errors were used. For multilevel models, cluster-robust variance estimation with small-sample corrections was used to account for the limited number of independent studies[40]. Preregistration of the overall analysis plan is available at https://osf.io/6t9vm.

We summarized the overall distribution of effects using a multilevel meta-analysis model, including fixed effects to distinguish replications from self-confirmatory tests, with random effects for each unique discovery and each unique ES nested within discovery[27]. The study-level variance component describes heterogeneity in the phenomena investigated in different studies and labs. The ES-level variance component describes heterogeneity across replications of the same phenomena.

Confirmation versus self-replication and independent replications. A random-effects meta-analysis was estimated to analyse the differences between the self-confirmatory test and the replication of the same discovery by the same lab. A negative average change would be evidence of declining replication ES, even when conducted by the same investigators.

Comparing self-confirmatory tests to replication results from other labs allows for assessment of the impact of between-lab differences in replicability success. Again, a random-effects meta-analysis was used to analyse differences between the ES in the self-confirmatory test and the average ES estimate in the three independent replications. Negative average differences would be evidence of declining replication ESs in cross-lab replication. The random-effects model provides an estimate of heterogeneity in the differences between self-confirmatory tests and replications beyond what would be expected by sampling error alone. Positive heterogeneity would indicate that ESs from self-confirmatory tests could not be exactly replicated by independent labs.

Slope across replications. According to one theory, declines in ESs over time are caused by a study being repeatedly run[25]. If accurate, the more studies run between the self-confirmatory test and the self-replication, the greater the decline. To examine temporal decline effects across all replications, we aggregated ES estimates from each self-confirmatory test with each of the replications and conducted a meta-analytic growth curve. The model also included random effects for each self-confirmatory test or replication attempt of each study that were allowed to covary within study according to an auto-regressive structure. The ESs were recoded for this analysis so that all effects were positive and a slope to non-significance or weakening ES would be negative in sign.

Reporting summary

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

Data availability

The data for each of the individual studies can be found by following the OSF links presented in supplementary information (section 3 and Supplementary Table 2). The data for the overall analysis presented here can be found at https://osf.io/bnq5j/.

Code availability

The statistical code for each of the individual studies can be found by following the OSF links presented in supplementary information (section 3 and Supplementary Table 2). The statistical code for the overall analysis presented here can be found at https://osf.io/bnq5j/.

References

1. Putnam, H. *Mathematics, Matter and Method: Volume 1, Philosophical Papers* (CUP Archive, 1975).
2. Schmidt, S. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* **13**, 90–100 (2009).
3. Begley, C. G. & Ellis, L. M. Drug development: raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
4. Camerer, C. F. et al. Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
5. 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).
6. Open Science Collaboration.Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
7. Errington, T. M. et al. Investigating the replicability of preclinical cancer biology. *eLife* **10**, e71601 (2021).
8. 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).
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. Wagenmakers, E. J. et al. Registered replication report: Strack, Martin, & Stepper (1988). *Perspect. Psychol. Sci.* **11**, 917–928 (2016).
11. O'Donnell, M. et al. Registered replication report: Dijksterhuis and van Knippenberg (1998). *Perspect. Psychol. Sci.* **13**, 268–294 (2018).
12. Piovesan, M., Døssing, F. S. & Wengström, E. R. Registered replication report: Rand, Greene & Nowak (2012). *Perspect. Psychol. Sci.* **12**, 527–542 (2017).
13. Verschuere, B. et al. Registered replication report on Mazar, Amir, and Ariely (2008). *Adv. Methods Pract. Psychol. Sci.* **1**, 299–317 (2018).
14. Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. Comment on 'Estimating the reproducibility of psychological science'. *Science* **351**, 1037 (2016).
15. Wilson, B. M., Harris, C. R. & Wixted, J. T. Science is not a signal detection problem. *Proc. Natl Acad. Sci. USA* **117**, 5559–5567 (2020).
16. Reich, E. S. Timing glitches dog neutrino claim. *Nat. News* **483**, 17 (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. Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J. & Reinero, D. A. Contextual sensitivity in scientific reproducibility. *Proc. Natl Acad. Sci. USA* **113**, 6454–6459 (2016).
19. Hedges, L. V. & Schauer, J. M. Consistency of effects is important in replication: rejoinder to Mathur and VanderWeele (2019). *Psychol. Methods* **24**, 576–577 (2019).
20. Glymour, C., Scheines, R. & Spirtes, P. *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling* (Academic Press, 2014).
21. Protzko, J. Null-hacking, a lurking problem. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/9y3mp (2018).
22. Bryan, C. J., Yeager, D. S. & O'Brien, J. Replicator degrees of freedom allow publication of misleading failures to replicate. *Proc. Natl Acad. Sci USA* **116**, 25535–25545 (2019).
23. Nosek, B. A. & Errington, T. M. What is replication? *PLoS Biol.* https://doi.org/10.1371/journal.pbio.3000691 (2020).
24. Schooler, J. Unpublished results hide the decline effect: some effects diminish when tests are repeated. *Nature* **470**, 437–438 (2011).
25. Protzko, J. & Schooler, J. W. Decline effects: types, mechanisms, and personal reflections. In *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions* (eds Lilienfeld, S. O. & Waldman, I. D.) 85–107 (Wiley Blackwell, 2017).
26. Yeager, D. S., Krosnick, J. A., Visser, P. S., Holbrook, A. L. & Tahk, A. M. Moderation of classic social psychological effects by demographics in the US adult population: new opportunities for theoretical advancement. *J. Pers. Soc. Psychol.* **117**, e84–e99 (2019).
27. Iso-Ahola, S. E. Reproducibility in psychological science: when do psychological phenomena exist? *Front. Psychol.* **8**, 879 (2017).
28. Iso-Ahola, S. E. Replication and the establishment of scientific truth. *Front. Psychol.* **11**, 2183 (2020).
29. Strack, F. & Stroebe, W. What have we learned? What can we learn? *Behav. Brain Sci.* **41**, e151 (2018).
30. Stroebe, W. & Strack, F. The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* **9**, 59–71 (2014).
31. Richard, F. D., Bond, C. F. Jr & Stokes-Zoota, J. J. One hundred years of social psychology quantitatively described. *Rev. Gen. Psychol.* **7**, 331–363 (2003).
32. Kvarven, A., Strømland, E. & Johannesson, M. Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-019-0787-z (2019).

33. Schweinsberg, M. et al. The pipeline project: pre-publication independent replications of a single laboratory's research pipeline. *J. Exp. Soc. Psychol.* **66**, 55–67 (2016).
34. Schooler, J. W. Metascience could rescue the 'replication crisis'. *Nature* **515**, 9 (2014).
35. Protzko, J. & Schooler, J. W. No relationship between researcher impact and replication effect: an analysis of five studies with 100 replications. *PeerJ* **8**, e8014 (2020).
36. 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).
37. Greenwald, A. G. Consequences of prejudice against the null hypothesis. *Psychol. Bull.* **82**, 1–20 (1975).
38. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**, 638–641 (1979).
39. Ernst, A. F., Hoekstra, R., Wagenmakers, E. J., Gelman, A. & van Ravenzwaaij, D. Do researchers anchor their beliefs on the outcome of an initial study? Testing the time-reversal heuristic. *Exp. Psychol.* **65**, 158–169 (2018).
40. Tipton, E. & Pustejovsky, J. E. Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *J. Educ. Behav. Stat.* **40**, 604–634 (2015).
41. Konstantopoulos, S. Fixed effects and variance components estimation in three-level meta-analysis. *Res. Synth. Methods* **2**, 61–76 (2011).
42. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48 (2010).
43. R Core Team. R: a language and environment for statistical computing. *R Foundation for Statistical Computing* https://www.R-project.org/ (2022).
44. Protzko, J., Buttrick, N., Ebersole, C. R., Lundmark, S. & Schooler, J. Not all explanations are equal: only explanations invoking a change from the true self mitigate responsibility. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/sx7cn (2020).
45. Protzko, J., Zedelius, C. M. & Schooler, J. W. Rushing to appear virtuous: time pressure increases socially desirable responding. *Psychol. Sci.* **30**, 1584–1591 (2019).
46. Evers, E. R. K., O'Donnell, M. & Inbar, Y. Arbitrary fairness in reward and punishments. *J. Exp. Psychol. Gen.* **152**, 956–967 (2023).
47. Laukkonen, R. E. et al. Irrelevant insights make worldviews ring true. *Sci. Rep.* **12**, 2075 (2022).
48. Inbar, Y. & Evers, E. Worse is bad: divergent inferences from logically equivalent comparisons. *J. Exp. Psychol. Gen.* **151**, 665–675 (2022).

Acknowledgements

We thank B. Fetzer and M. Kidwell for their help with the project. We also thank B. Ouimette, P. Noel, B. Schembari, M. Gross, B. Kaveladze, R. Laukkonen, N. Buchen, R. Narr, M. Austin, J. Rodriguez, K. Kurosawa, H. Harrington, M. Basili and D. Dienhart for their help in completing the individual studies comprising this project. We thank S. Hoogeveen, A. Sarafoglou, A. Dreber Almenberg, and E. J. Wagenmakers for help in creating the study descriptions. This work was sponsored by grants from the Fetzer Franklin Fund of the John E. Fetzer Memorial Trust to each of the participating laboratories at the University of California, Santa Barbara; the University of Virginia; the University of California, Berkeley; and Stanford University. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript, aside from the participation of J.W.

Author contributions

Conceptualization: J.P., J.K., L.N., B.A.N., J.W. and J.W.S. Methodology: J.P., J.K., M.B., S.L., J.E.P., J.W. and J.W.S. Validation: J.P., J.K., L.N., B.A.N., J.A., M.B., N.B., M.D., C.R.E., S.L., B.M., M.O., H.P. and S.S.R. Formal analysis: J.P., M.D. and J.E.P. Investigation: J.P., J.K., L.N., B.A.N., J.A., M.B., N.B., M.D., C.R.E., S.L., B.M., M.O., H.P., S.S.R. and J.W.S. Resources: J.P., L.N., B.A.N., J.A., M.B., N.B., M.D., C.R.E., S.L., B.M., M.O., H.P. and S.S.R. Software: S.L. Data curation: J.P., J.K., M.B., S.L. and J.E.P. Writing—original draft: J.P. and J.W.S. Writing—review and editing: J.P., J.K., L.N., B.A.N., J.A., M.B., N.B., M.D., C.R.E., S.L., B.M., M.O., H.P., J.E.P., J.W. and J.W.S. Project administration: J.P., J.W. and J.W.S. Supervision: J.K., L.N., B.A.N., M.B., J.W. and J.W.S. Funding acquisition: J.K., L.N., B.A.N., J.W. and J.W.S. Visualization: B.A.N. and J.E.P. Author names are presented in the following order: project manager (J.P.), individual principal investigators in alphabetical order (J.K., L.N. and B.A.N.), authors in alphabetical order (J.A., M.B., N.B., M.D., C.R.E., S.L., B.M., M.O., H.P., J.E.P., S.S.R. and J.W.), and project principal investigator and senior author (J.W.S.).

Competing interests

B.A.N. is the executive director of the non-profit Center for Open Science with a mission to "increase openness, integrity, and reproducibility of research". J.W. was the scientific director of the Fetzer Franklin Fund that sponsored this research, and B.A.N. and J.W.S were also on the scientific advisory board. J.W. made substantive contributions to the design and execution of this research, but as a funder did not have controlling interest in the decision to publish or not. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-023-01749-9.

Correspondence and requests for materials should be addressed to John Protzko.

Peer review information *Nature Human Behaviour* thanks Malcolm Macleod and Malte Elson for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

nature portfolio

Corresponding author(s): John Protzko

Last updated by author(s): Jul 21, 2023

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	All analyses were conducted using the R statistical computing environment (Version 4.2.1). All analyses used meta-analytic random effects models estimated using restricted maximum likelihood with the metafor package (Version 2.1.0). Standard errors and confidence intervals for all analyses were calculated using cluster-robust standard errors (CR2-type), clustering by study, using the clubSandwich package (Version 0.3.5).
Data analysis	All analyses were conducted using the R statistical computing environment (Version 4.2.1). All analyses used meta-analytic random effects models estimated using restricted maximum likelihood with the metafor package (Version 2.1.0). Standard errors and confidence intervals for all analyses were calculated using cluster-robust standard errors (CR2-type), clustering by study, using the clubSandwich package (Version 0.3.5).

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 for each of the individual studies can be found following the OSF links presented in Table S1 in the SI. Data for the overall analysis presented here can be found at https://osf.io/bnq5j/.

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, the data here is a meta-analysis and no effect sizes were reported broken down by sex or gender.
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	This is a meta-analysis of studies. The complete population of studies is included.
Recruitment	Each study in this prospective meta-analysis was put forward for replication, and the confirmation and all replications are included in this meta-analysis.
Ethics oversight	The individual studies in this meta-analysis were all approved by the IRBs at University of California, Santa Barbara, Stanford University, University of Virginia, and University of California, Berkeley. This current study is a meta-analysis and does not involve human subjects and is exempt. The prediction survey of lay participants was found exempt by the Office for Research on Human Subjects at the University of California, Santa Barbara.

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	A prospective meta-analysis of quantitative effect sizes from new discoveries and their replications.
Research sample	The sample here are effect sizes from prospective meta-analysis of new discoveries and replications by the four labs. There is no publication bias in this sample as the full population of studies is used.
Sampling strategy	Each lab put forward a study for replication, and it was replicated by all labs. Each study was powered at N = 1,500 to be able to detect very small effect sizes. Based on a power analysis of the 13 self-confirmatory tests with statistically-significant results, the average replication power was 0.96 with a median approaching 1 and average power in replication studies of specific discoveries ranging from 0.62 to approaching 1.
Data collection	All labs studies effect sizes were calculated by them and verified by one of the teams. All data was then put into a .csv file that was used for data analysis.
Timing	Studies started 06/2016 and ended 07/2019, survey of participants occurred 12/2022
Data exclusions	No study put forward was excluded.
Non-participation	The survey of particpants was done on Prolific academic, non-participation is not possible to determine.
Randomization	Order of studies and replications were randomized in a latin square design.

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