nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-018-0399-z

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer[1,16], Anna Dreber[2,16], Felix Holzmeister[3,16], Teck-Hua Ho[4,16], Jürgen Huber[3,16], Magnus Johannesson[2,16], Michael Kirchler[3,5,16], Gideon Nave[6,16], Brian A. Nosek[7,8,16]*, Thomas Pfeiffer[9,16], Adam Altmejd[2], Nick Buttrick[7,8], Taizan Chan[10], Yiling Chen[11], Eskil Forsell[12], Anup Gampa[7,8], Emma Heikensten[2], Lily Hummer[8], Taisuke Imai[13], Siri Isaksson[2], Dylan Manfredi[6], Julia Rose[3], Eric-Jan Wagenmakers[14] and Hang Wu[15]

Being able to replicate scientific findings is crucial for scientific progress[1–15]. We replicate 21 systematically selected experimental studies in the social sciences published in *Nature* and *Science* between 2010 and 2015[16–36]. The replications follow analysis plans reviewed by the original authors and pre-registered prior to the replications. The replications are high powered, with sample sizes on average about five times higher than in the original studies. We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size. Replicability varies between 12 (57%) and 14 (67%) studies for complementary replicability indicators. Consistent with these results, the estimated true-positive rate is 67% in a Bayesian analysis. The relative effect size of true positives is estimated to be 71%, suggesting that both false positives and inflated effect sizes of true positives contribute to imperfect reproducibility. Furthermore, we find that peer beliefs of replicability are strongly related to replicability, suggesting that the research community could predict which results would replicate and that failures to replicate were not the result of chance alone.

To what extent can we trust scientific findings? The answer to this question is of fundamental importance[1–3], and the reproducibility of published studies has been questioned in many fields[4–10]. Until recently, systematic evidence has been scarce[11–15]. The Reproducibility Project: Psychology (RPP)[12] put the question of scientific reproducibility at the forefront of scientific debate[37–39]. The RPP replicated 100 original studies in psychology and found a significant effect in the same direction as the original studies for 36% of the 97 studies reporting 'positive findings'[12]. The RPP was followed by the Experimental Economics Replication Project (EERP), which replicated 18 laboratory experiments in economics and found a significant effect in the same direction as the original studies for 61% of replications[13]. Both the RPP and the EERP had high statistical power to detect the effect sizes observed in the original studies. However, the effect sizes of published studies may be inflated even for true-positive findings owing to publication or reporting biases[40–42]. As a consequence, if replications were well powered to detect effect sizes smaller than those observed in the original studies, replication rates might be higher than those estimated in the RPP and the EERP.

We provide evidence about the replicability of experimental studies in the social sciences published in the two most prestigious general science journals, *Nature* and *Science* (the Social Sciences Replication Project (SSRP)). Articles published in these journals are considered exciting, innovative and important. We include all experimental studies published between 2010 and 2015 that (1) test for an experimental treatment effect between or within subjects, (2) test at least one clear hypothesis with a statistically significant finding, and (3) were performed on students or other accessible subject pools. Twenty-one studies were identified to meet these criteria. We used the following three criteria in descending order to determine which treatment effect to replicate within each of these 21 papers: (a) select the first study reporting a significant treatment effect for papers reporting more than one study, (b) from that study, select the statistically significant result identified in the original study as the most important result among all within- and between-subject treatment comparisons, and (c) if there was more than one equally central result, randomly select one of them for replication. The interpretation of which was the most central and important statistically significant result within a study in criteria (b) above was made by us and not by the original authors. See Supplementary Methods and Supplementary Tables 1 and 2 for details.

[1]California Institute of Technology, Pasadena, CA, USA. [2]Department of Economics, Stockholm School of Economics, Stockholm, Sweden. [3]Department of Banking and Finance, University of Innsbruck, Innsbruck, Austria. [4]NUS Business School, National University of Singapore, Singapore, Singapore. [5]Centre for Finance, Department of Economics, University of Göteborg, Göteborg, Sweden. [6]The Wharton School, University of Pennsylvania, Philadelphia, PA, USA. [7]Department of Psychology, University of Virginia, Charlottesville, VA, USA. [8]Center for Open Science, Charlottesville, VA, USA. [9]New Zealand Institute for Advanced Study, Auckland, New Zealand. [10]Office of the Senior Deputy President and Provost, National University of Singapore, Singapore, Singapore. [11]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [12]Spotify Sweden AB, Stockholm, Sweden. [13]Department of Economics, LMU Munich, Munich, Germany. [14]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands. [15]School of Management, Harbin Institute of Technology, Harbin, China. [16]These authors contributed equally: Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer. *e-mail: nosek@cos.io

To address the possibility of inflated effect sizes in the original studies, we used a high-powered design and a two-stage procedure for conducting the replications. In stage 1, we had 90% power to detect 75% of the original effect size at the 5% significance level in a two-sided test. If the original result replicated in stage 1 (a two-sided $P<0.05$ and an effect in the same direction as in the original study), no further data collection was carried out. If the original result did not replicate in stage 1, we carried out a second data collection in stage 2 to have 90% power to detect 50% of the original effect size for the first and second data collections pooled.

The motivation for having 90% power to detect 50% of the original effect size was based on the replication effect sizes in the RPP being on average about 50% of the original effect sizes[12] (see Supplementary Methods for details; the average relative effect size of the replications in the EERP was 66%[13]). On average, replication sample sizes in stage 1 were about three times as large as the original sample sizes and replication sample sizes in stage 2 were about six times as large as the original sample sizes. All of the replication and analysis plans were made publicly known on the project website, pre-registered at the Open Science Framework (OSF) and sent to the original authors for feedback and verification prior to data collection (the pre-replication versions of the replication reports and the final versions are posted at the project's OSF repository (https://osf.io/pfdyw/); the final versions of the replication reports include a section called 'Unplanned protocol deviations', which lists any deviations from the pre-registered replication protocols and these deviations are also listed towards the end of the Supplementary Methods. There was no deviation from the protocol for 7 replications[17,18,20–22,25,35], minor deviations for 12 replications[19,23,24,26,27,29–34,36], an unintended methodological deviation for one replication[28], and a continuation to the stage 2 data collection by mistake for one replication[16].

There is no universally agreed on criterion for replication[12,43–46], but our power analysis strategy is based on detecting a significant effect in the same direction as the original study using the same statistical test. As such, we treat this as the primary indicator of replication and refer to it as the statistical significance criterion. This approach is appealing for its simplicity as a binary measure of replication, but does not fully represent evidence of reproducibility. We also provide results for the relative effect size of the replication as a continuous measure of the degree of replication. To complement these indicators, we present results for: (1) a meta-analytic estimate of the original and the replication results combined[12], (2) 95% prediction intervals[47], (3) the 'small telescopes' approach[46], (4) the one-sided default Bayes factor[48], (5) a Bayesian mixture model[49], and (6) peer beliefs about replicability[50]. See Supplementary Methods and Supplementary Figs. 1–3 for additional robustness tests of the replication results.

In stage 1, we find a significant effect in the same direction as the original study for 12 replications[16–19,22–25,27,29,30,36] (57.1%) (Fig. 1a and Supplementary Table 3). When we increase the statistical power further in stage 2 (Fig. 1b and Supplementary Table 4), two additional studies[20,31] replicate based on this criterion. By mistake, a second data collection was carried out for one study[16] replicating in stage 1; thus, we also include this study in the stage 2 results to base our results on all the data collected. This study[16] does not replicate in stage 2. This may suggest that replication studies should routinely be powered to detect at least 50% of the original effect size or that one should use a lower P value threshold than 0.05 for not continuing to stage 2 in our two-stage testing procedure. Based on all of the data collected, 13 (61.9%) studies replicated after stage 2 using the statistical significance criterion.

The mean standardized effect size (correlation coefficient r) of the replications is 0.249, compared to 0.460 in the original studies (Supplementary Fig. 4). This difference is significant (Wilcoxon signed-ranks test, $z=3.667$, $P<0.001$, $n=21$) and the mean relative effect size of the replications is 46.2%. For the 13 studies that replicated, the mean relative effect size is 74.5%, and for the 8 studies that did not replicate, the mean relative effect size is 0.3%. It is not surprising that the mean relative effect size is smaller for the non-replicating effects than for the replicating effects as these are correlated indicators. However, it is notable that, even among the replicating effects, the effect sizes for the replications were weaker than the original findings, and for the non-replicating effects, the mean effect sizes were approximately zero.

We also combined the original result and the replication in a meta-analytic estimate of the effect size. As seen in Fig. 1c, 16 studies (76.2%) have a significant effect in the same direction as the original study in the meta-analysis. However, the meta-analysis assumes that the results of the original studies are not influenced by publication or reporting biases, making the meta-analytic results an overly optimistic indicator compared to criteria that focused on the replication evidence[12]. A team recently suggested that the P value threshold for significant findings should be lowered from 0.05 to 0.005 for new discoveries[51]. In a replication context, it would be relevant to apply this stricter threshold to meta-analytic results. In this case, the meta-analysis leads to the same conclusions about replication as our primary replication indicator (that is, 13 studies or 61.9% of studies have a $P<0.005$ in the meta-analysis). It is obvious that the 13 successful replications would achieve $P<0.005$ when the original and replication results were pooled, but this criterion could have also included replications that did not achieve $P<0.05$ but were in the right direction and were combined with an original study with particularly strong evidence.

A complementary replication criterion is to count how many replicated effects lie in a 95% prediction interval[47], which takes into account the variability in both the original study and the replication study. Using this method, 14 effects replicated (66.7%; see Fig. 2a and Supplementary Methods for details). This method yields the same replication outcome as the statistical significance criterion for 20 of the 21 studies.

The small telescopes approach estimates whether the replication effect size is significantly smaller than a 'small effect' in the original study with a one-sided test at the 5% level. A small effect is defined as the effect size that the original study would have had 33% power to detect. Following the small telescopes approach[46], 12 studies (57.1%) replicate (see Fig. 2b and Supplementary Methods for details). One replication has a significant effect in the same direction as the original study, but the effect size is significantly smaller than a small effect as defined by the small telescopes approach. This is the only difference compared to the statistical significance criterion.

Another way to represent the strength of evidence in favour of the original result versus the null hypothesis of no effect is to estimate the Bayes factor[45,48,52,53]. The Bayes factor compares the predictive performance of the null hypothesis against that of an alternative hypothesis in which the uncertainty about the true effect size is quantified by a prior distribution. The prior distributions were first set to their generic defaults; they were then folded across the test value so that all prior mass was consistent with the direction of the effect from the original study, thereby implementing a Bayesian one-sided test (see the Supplementary Methods for details). For example, the replication of Pyc and Rawson[31] yielded a one-sided default Bayes factor of $\mathrm{BF}_{+0}=6.8$, meaning that the one-sided alternative hypothesis out predicted the null hypothesis of no effect by a factor of almost 7.

The one-sided default Bayes factor exceeds 1, providing evidence in favour of an effect in the direction of the original study for the 13 (61.9%) studies that replicated according to our primary replication indicator (Fig. 3). This evidence is strong to extreme for 9 (42.9%) studies. The default Bayes factor is below 1 for 8 (38.1%) studies, providing evidence in support of the null hypothesis; this evidence is strong to extreme for 4 (19.0%) studies.

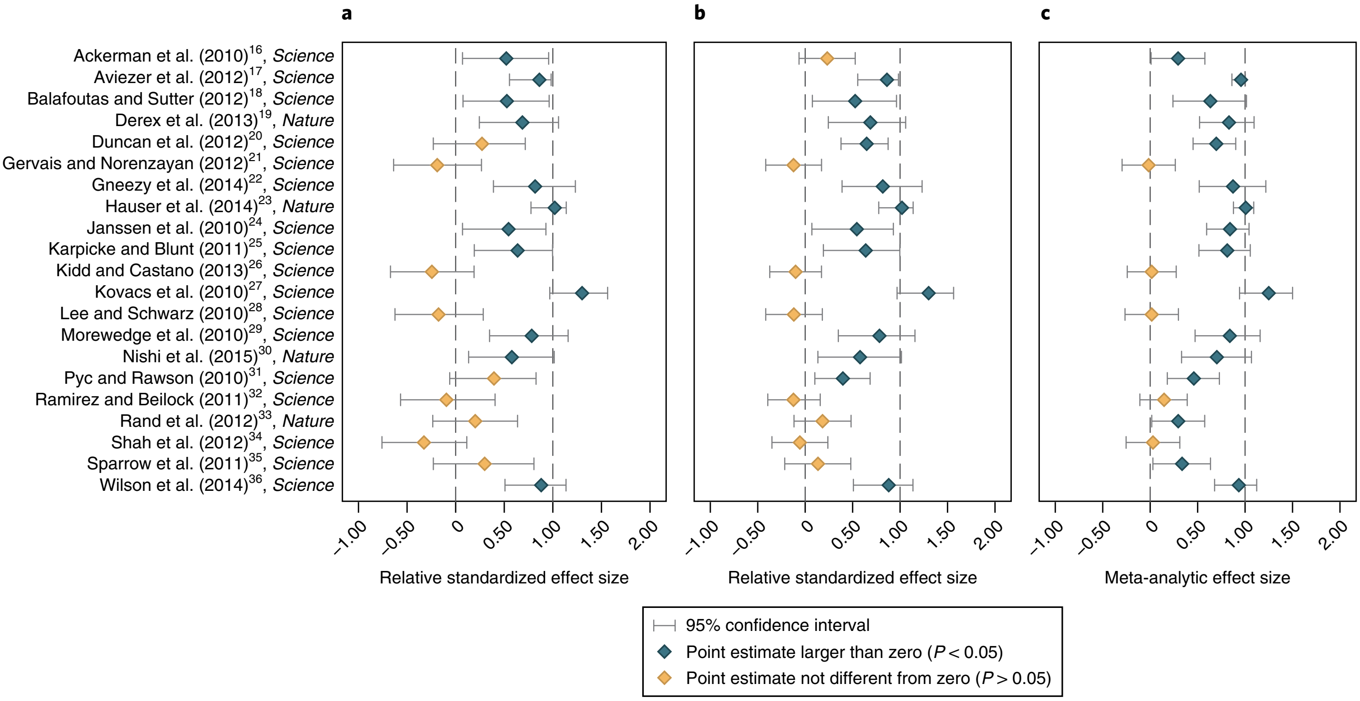

Fig. 1 | Replication results after stage 1 and stage 2. a, Plotted are the 95% CIs of the replication effect sizes (standardized to the correlation coefficients *r*) after stage 1. The standardized effect sizes are normalized so that 1 equals the original effect size. There is a significant effect in the same direction as in the original study for 12 out of 21 replications (57.1%; 95% CI = 34.1–80.2%). **b**, Plotted are 95% CIs of replication effect sizes (standardized to the correlation coefficients *r*) after stage 2 (replications not proceeding to stage 2 are included with their stage 1 results). The standardized effect sizes are normalized so that 1 equals the original effect size. There is a significant effect in the same direction as in the original study for 13 out of 21 replications (61.9%; 95% CI = 39.3–84.6%). **c**, Meta-analytic estimates of effect sizes combining the original and the replication studies. Shown are the 95% CIs of the standardized effect sizes (correlation coefficient *r*). The standardized effect sizes are normalized so that 1 equals the original effect size. Original and zero effect size are indicated by dashed lines. Sixteen out of 21 studies have a significant effect in the same direction as the original study in the meta-analysis (76.2%; 95% CI = 56.3–96.1%). Any deviations from the pre-registered replication protocols are listed towards the end of the Supplementary Methods. There was no deviation from the protocol for 7 replications[17,18,20–22,25,35], minor deviations for 12 replications[19,23,24,26,27,29–34,36], an unintended methodological deviation for one replication[28] and a continuation to the stage 2 data collection by mistake for one replication[16].

In additional Bayesian analyses, we use an errors-in-variables mixture model[49] to estimate the true-positive rate in the total sample (see the Supplementary Methods and Supplementary Fig. 5 for details). The estimated true-positive rate is 67% (Supplementary Fig. 5), which is close to the other replicability estimates. The mixture model also estimates that the average relative effect size of true positives is 71% (Supplementary Fig. 5), suggesting that the original studies overestimated the effect sizes of true positives.

We also estimate peer beliefs about replicability using surveys and prediction markets[50,54] (see Supplementary Methods, Supplementary Table 5 and Supplementary Fig. 6 for details). The prediction markets produce a collective peer estimate of the probability of replication that can be interpreted as a reproducibility indicator[50]. The average prediction market belief of replicating after stage 2 is a replication rate of 63.4% and the average survey belief is 60.6%, which are both close to the observed replication rate of 61.9% (Fig. 4; see Supplementary Methods, Supplementary Figs. 7 and 8 and Supplementary Tables 5 and 6 for more details). The prediction market beliefs and the survey beliefs are highly correlated and both are highly correlated with a successful replication (Fig. 4 and Supplementary Fig. 7); that is, in the aggregate, peers were very effective at predicting future replication success.

In the RPP[12] and the EERP[13], replication success was negatively correlated with the *P* value of the original study, suggesting that original study *P* values might be a predictor of replicability. We also find a negative correlation between the *P* value of the original study and replication success, although it is not significant (Spearman correlation coefficient: −0.405, $P = 0.069$, 95% CI = −0.712 to 0.033, $n = 21$); the estimate is in between the correlations found in the RPP (−0.327) and the EERP (−0.572) (Supplementary Table 7). That peers are to some extent able to predict which studies are most likely to replicate suggests that there are features of the original studies that journals or researchers can use in determining ex ante whether a study is likely to replicate. Taken together, the results from the RPP, EERP and SSRP suggest that the *P* value of the original study is one such important determinant of replication. The SSRP with $n = 21$ studies is too small to reliably test determinants of replications, but pooling the results of all large-scale replication projects may offer a higher-powered opportunity to explore moderators of replication.

To summarize, we successfully replicated 13 out of 21 findings from experimental social and behavioural science studies published in *Science* or *Nature* between 2010 and 2015 based on the statistical significance criterion with very high-powered studies compared to the RPP[12] and the EERP[13]. This number is larger than the replication rate of the RPP and similar to the replication rate of the EERP (Supplementary Fig. 9). However, the small sample of studies and different selection criteria make it difficult to draw any interpretation confidently in comparison with those studies. However, we can conclude that increasing power substantially is not sufficient to reproduce all published studies. Furthermore, we observe that the conclusions across binary replication criteria converge with increased statistical power. The small telescopes and the 95% prediction interval indicators drew different conclusions on only one of the replications compared to the statistical significance criterion.

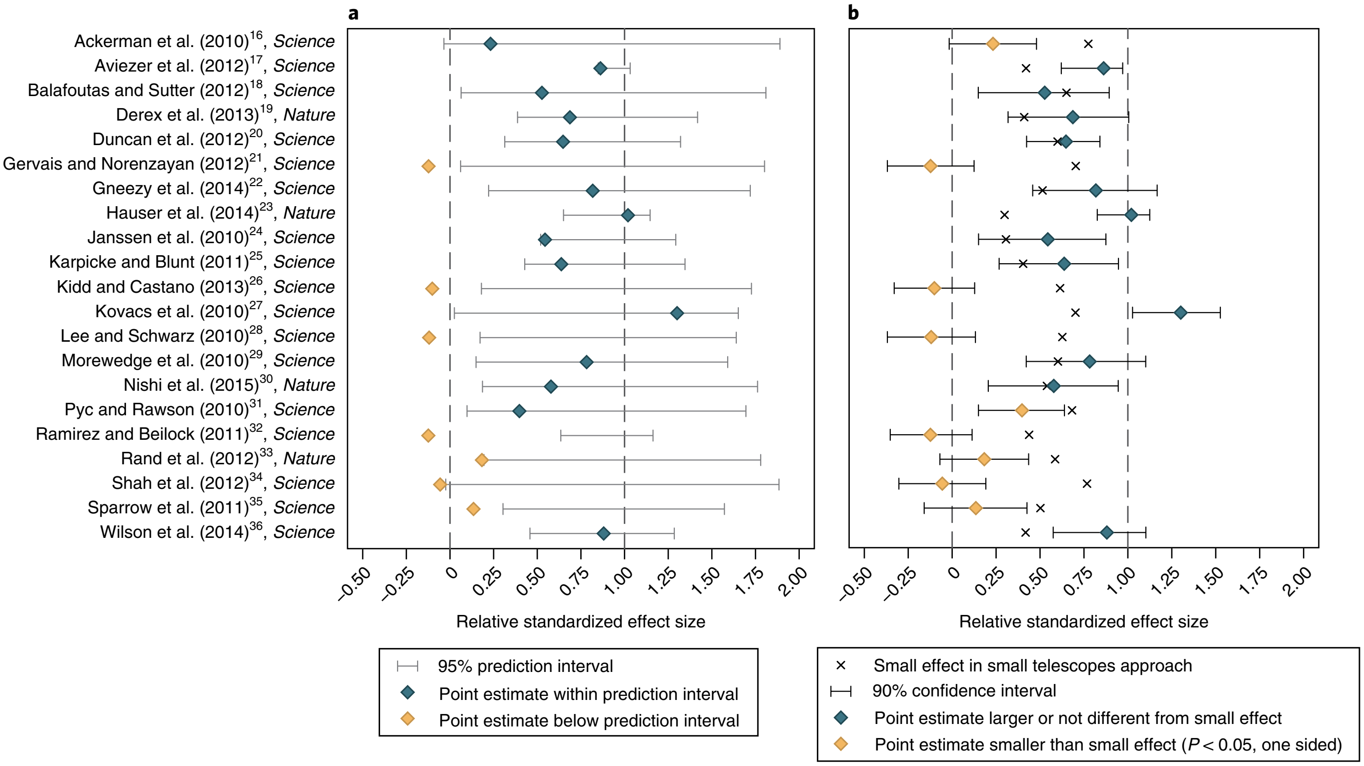

Fig. 2 | Replication results for two complementary replication indicators. a, Plotted are the 95% prediction intervals[47] for the standardized original effect sizes (correlation coefficient *r*). The standardized effect sizes are normalized so that 1 equals the original effect size. Original and zero effect size are indicated by dashed lines. Fourteen replications out of 21 (66.7%; 95% CI = 44.7–88.7%) are within the 95% prediction interval and replicate according to this indicator. **b**, Plotted are the 90% CIs of replication effect sizes in relation to small-effect sizes as defined by the small telescopes approach[46] (the effect size that the original study would have had 33% power to detect). Effect sizes are standardized to correlation coefficients *r* and normalized so that 1 equals the original effect size. A study is defined as failing to replicate if the 90% CI is below the small effect. According to the small telescopes approach, 12 out of 21 (57.1%; 95% CI = 34.1–80.2%) studies replicate.

Considering statistical significance and effect sizes simultaneously, we observe two major outcomes. First, even among successful replications, the estimated effect sizes were smaller than the original study. For the 13 studies that replicated according to the statistical significance criterion, the replication effect sizes were about 75% of the original effect size. This provides an estimate of the overestimation of effect sizes of true positives in the original studies. The Bayesian mixture model corroborates this result, yielding an estimate of the relative effect size of true positives of 71%. This implies that meta-analyses of true-positive findings will overestimate effect sizes on average. This finding bolsters evidence that the existing literature contains exaggerated effect sizes because of pervasive low-powered research coupled with bias selecting for significant results for publication[8,12]. In addition, if this finding generalizes to the literatures investigated by the RPP and the EERP, it suggests that the statistical power of these two projects, in which the sample sizes were determined to obtain 90% power to detect the original effect size, was de-facto smaller than intended. This would imply that the replication rates, based on the statistical significance criterion, were underestimated in these studies, consistent with those authors' speculation.

Second, among the unsuccessful replications, there was essentially no evidence for the original finding. The average relative effect size was very close to zero for the eight findings that failed to replicate according to the statistical significance criterion. The expected relative effect size for a sample of false positives is zero, but this observation does not demand the conclusion that the eight original findings were false positives. Another possibility is that the replication studies failed to implement necessary features of the protocol to detect the effect[38]. We cannot rule out this alternative, but we also do not have evidence for necessary features missing from the replications that would reduce the observed effect sizes to zero. Indeed, it would be surprising but interesting to identify an unintended difference that completely eliminated the effect rather than just reduce the effect size. One suggested indicator for whether differences between studies are a likely cause for bias is the endorsement of the original authors[38]. In the current project, we took extensive efforts to ensure that the replications would be as close as possible to the originals. All of the replications but one[35] were designed with the collaboration of the original authors (for the replication[35] that was not designed with the collaboration of the original authors, the original authors did not respond to our queries). Furthermore, all of the reviewed replications but one[32] were approved by the original authors. However, none of this implies that the original authors agree with the final outcomes or interpretation. For example, changes in planned implementation or insights after observing the results could lead to different interpretations of the replication outcome and ideas for subsequent research to clarify the understanding of the phenomenon. See the Supplementary Methods and the posted replication reports for each study for more details, including follow-up comments from the original authors if provided. For more information, see the Correspondences by the original authors published alongside this Letter (Duncan and Davachi; Gervais and Norenzayan; Kidd and Castano; Lee and Schwarz; Pyc and Rawson; Rand; Shah et al.; and Sparrow).

Another hypothesis that could account for replication failures, at least partly, is the result of chance, such as a large degree of heterogeneity in treatment effects in different samples[38]. However, such

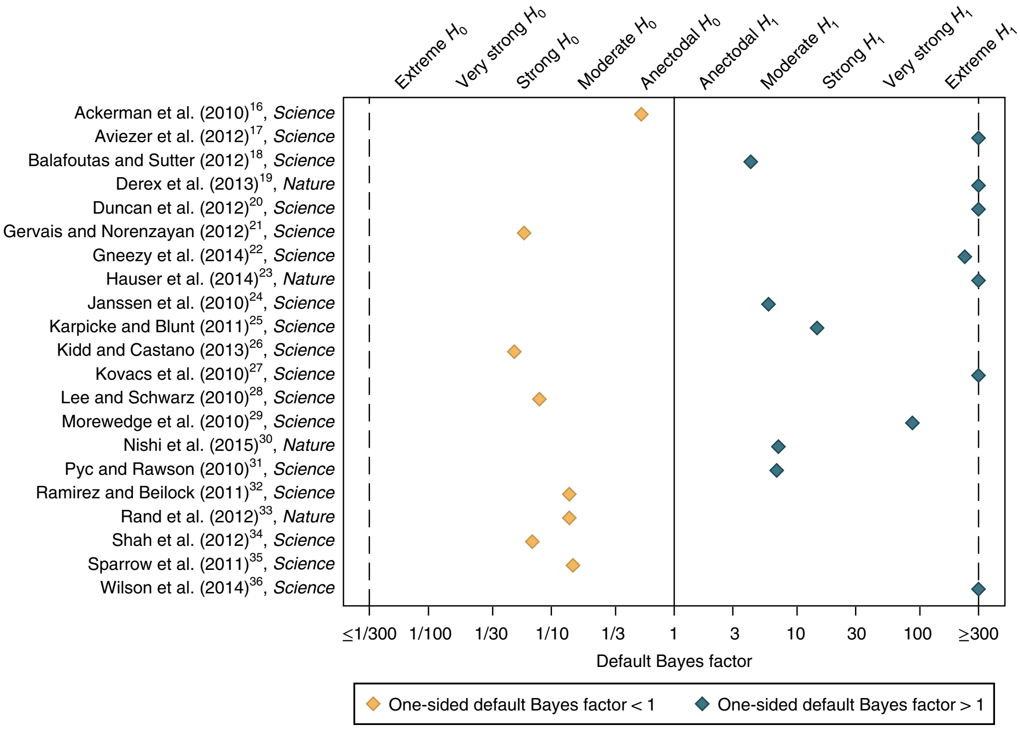

Fig. 3 | Default Bayes factors (one sided) for the 21 replications. A default Bayes factor[48] above 1 favours the hypothesis of an effect in the direction of the original paper and a default Bayes factor below 1 favours the null hypothesis (H_0) of no effect. The evidence categories proposed by Jeffreys[52] are also shown (from extreme support for the null hypothesis to extreme support for the original hypothesis). The default Bayes factor is above 1 and provides evidence in favour of an effect in the direction of the original study for the 13 out of 21 (61.9%) studies that replicated according to the statistical significance criterion. This evidence is strong to extreme for 9 out of 21 (42.9%) studies. The default Bayes factor is below 1 for 8 out of 21 (38.1%) studies, providing evidence in support of the null hypothesis; this evidence is strong to extreme for 4 out of 21 (19.0%) studies. Values more extreme than 1/300 or 300 are represented on the dashed lines. H_1, alternative hypothesis.

heterogeneity would not affect the average relative effect size of replications, as replications would be as likely to overestimate as underestimate the original effect sizes. Thus, it cannot explain why the average effect sizes of our replications is only about 50% of the original effect sizes. Furthermore, the strong correlation between the peer predictions and the observed replicability is discordant with the possibility that replication failures occurred by chance alone. That is, researchers seem to have identified a priori systematic differences between the studies that replicated and those that did not. This capacity to predict the replicability of effects is a reason for optimism that methods will emerge to anticipate reproducibility challenges and guide efficient use of replication resources towards exciting but uncertain findings.

Below, we discuss some limitations of the SSRP. The SSRP is a small sample of studies with specific selection criteria for experimental studies from two high-profile journals. Work that is published in *Nature* and *Science* may be atypical to the field as a whole and may have a stronger focus on novelty, which may also lead to greater—or lesser—editorial scrutiny. The small sample and selective criteria significantly reduce confidence in generalizing these findings to the social science literature more generally. Indeed, like all other research, replications require an accumulation of evidence across multiple efforts to identify and address sampling biases and to obtain increasingly precise estimates of replicability. This study adds to this accumulating literature with a focused, high-powered investigation of high-profile studies published in *Nature* and *Science*. Notably, with replication sample sizes about five times larger as the original studies, we get relatively precise estimates of the individual effects of these single replications and the average relative effect sizes that are very similar to what was observed in RPP.

Another important limitation is that, for papers reporting a series of studies, we only replicate one of those studies, and for studies testing more than one hypothesis, we only replicate one hypothesis. Like previous large-scale replication projects, this study does not provide definitive insight on any of the original papers from which we designed the replication studies. An alternative methodology would be to replicate all results within the selected study or all results within all studies in a paper reporting a series of studies. This would give more information from each replication and a more precise estimate of reproducibility of each study and paper. All investigations involve trade-offs. The advantage of an in-depth examination of a hypothesis within a study is greater insight and precision of the reproducibility of its findings. The disadvantage is that many fewer findings can be investigated to learn about the reproducibility of findings more generally. Some other findings reported in the original papers can be tested with the data available in the replications of our study. We did not consider those secondary findings in this paper or in deciding the statistical power plans for the design. However, all of our data and materials are publicly posted on OSF and will be available to other researchers who may want to pursue this issue further in follow-up work.

The original authors in reviewing our paper and replication results have noted some limitations on the replications of their individual studies. These are discussed more in the Supplementary Information; several of the original authors have also posted comments on the replications at the OSF alongside our replication reports. For example, previously unidentified or inadvertent changes to the protocol may have affected replication success for some studies. For more information, see also the Correspondences by the original authors published alongside this Letter. In addition,

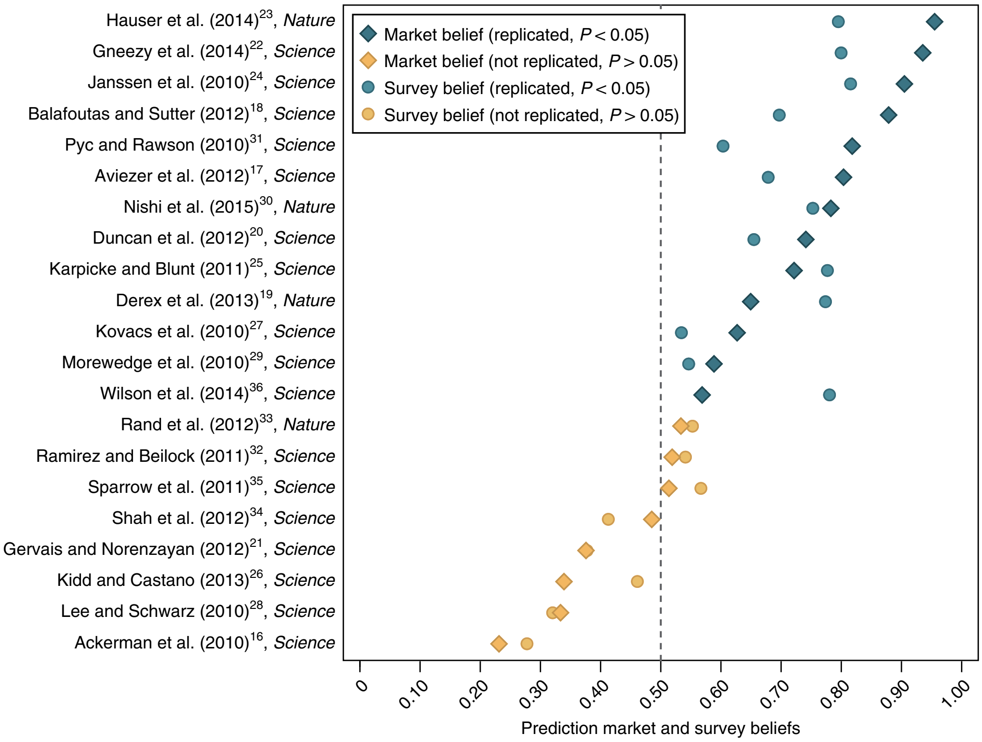

Fig. 4 | Prediction market and survey beliefs. The prediction market beliefs and the survey beliefs of replicating (from treatment 2 for measuring beliefs; see the Supplementary Methods for details and Supplementary Fig. 6 for the results from treatment 1) are shown. The replication studies are ranked in terms of prediction market beliefs on the *y* axis, with replication studies more likely to replicate than not to the right of the dashed line. The mean prediction market belief of replication is 63.4% (range: 23.1–95.5%, 95% CI = 53.7–73.0%) and the mean survey belief is 60.6% (range: 27.8–81.5%, 95% CI = 53.0–68.2%). This is similar to the actual replication rate of 61.9%. The prediction market beliefs and survey beliefs are highly correlated, but imprecisely estimated (Spearman correlation coefficient: 0.845, 95% CI = 0.652–0.936, $P < 0.001$, $n = 21$). Both the prediction market beliefs (Spearman correlation coefficient: 0.842, 95% CI = 0.645–0.934, $P < 0.001$, $n = 21$) and the survey beliefs (Spearman correlation coefficient: 0.761, 95% CI = 0.491–0.898, $P < 0.001$, $n = 21$) are also highly correlated with a successful replication.

for papers reporting a series of studies, we replicated the first study that reported a significant treatment effect. In some cases, the original authors argue that other studies in their papers report more important results or use stronger research designs[26,34] (see the Correspondence by Kidd and Castano, and the Correspondence by Shah et al.). If the replicability of the first study systematically differs from the replicability of subsequent studies in a paper, our criteria for deciding which study to replicate will systematically overestimate or underestimate replicability.

Inspired by our replication, the original authors of Shah et al.[34] decided to carry out a replication study of their own on all five of their studies (with results posted at https:osf.io/vzm23/). They did replicate what they consider to be their most important finding: scarcity itself leads to overborrowing. They also failed to replicate study 1 in their paper, consistent with our findings. Their approach of conducting replications of their own studies is admirable and provides additional insight and precision for understanding those effects.

Five of our replications were carried out on Amazon Mechanical Turk (AMT), and for one of those (Rand et al.[33]), the original authors argue that increasing familiarity with economic game paradigms among AMT samples may have decreased the replicability of their result (see the Correspondence by Rand). It cannot be ruled out that changes in the AMT subject pool over time have affected our results, but we also note that the two other studies based on economic game paradigms and AMT data replicated successfully[23,50]. It would be interesting in future work to test whether replicability differs for older versus newer studies or depends on the time that has elapsed between the original study and the replication.

For the Sparrow et al.[35] replication, the original authors did not provide us with any materials for the replication or feedback on our inquiries. This made it more difficult to replicate the experimental design of the original study. After the replication had been completed, the original authors noted some design differences compared to the original study (see the Correspondence by Sparrow). These design differences are discussed further in the Supplementary Information and we cannot rule out that they influenced the replication result. This illustrates the importance of open access to all of the materials of published studies for conducting direct replications and accumulating scientific knowledge.

The observed replication rate of 62%, based on the statistical significance criterion, adds to a growing pool of replicability rates from various systematic replication efforts with distinct selection and design criteria: the RPP[12] (36%, $n = 100$ studies), the EERP[13] (61%, $n = 18$ studies), Many Labs 1[11] (77%; $n = 13$ studies), Many Labs 2[15] (50%, $n = 28$ studies) and Many Labs 3[14] (30%, $n = 10$ studies). It is too early to draw a specific conclusion about the reproducibility rates of experimental studies in the social and behavioural sciences. Each investigation has a relatively small sample of studies with idiosyncratic inclusion criteria and unknown generalizability. However, the diversity in approaches provides some confidence that considering them in the aggregate may provide more general insight about reproducibility in the social behavioural sciences. As a descriptive and speculative interpretation of these findings in the aggregate, we believe that reasonable lower-bound and upper-bound estimates are 35% and 75%, respectively, for an average reproducibility rate of published findings in social and behavioural sciences. Accumulating additional evidence will reveal whether there are systematic biases in these reproducibility estimates themselves.

When assessing reproducibility, we are interested in both the systematic bias in the estimated effect sizes of the original studies and the fraction of original hypotheses that are directionally true. The average relative effect size of 50% in the SSRP is a direct estimate of the systematic bias in the published findings of the 21 studies, as it should be 100% if the original studies provide unbiased estimates of true-effect sizes. This estimate assumes that there is no systematic difference in the effectiveness of implementing the study procedures or the appropriateness of testing circumstances between the original and the replication studies. If both of those assumptions are true, then our data indicate that the systematic bias is partly due to false positives and partly due to the overestimated effect sizes of true positives. These systematic biases can be reduced by implementing pre-registration of analysis plans to reduce the likelihood of false positives and registration and reporting of all study results to reduce the effects of publication bias inflating effect sizes[55]. With notable progress on these practices, particularly in the social and behavioural sciences[56], we predict that replicability will improve over time.

Methods

The methods of the study are detailed in the Supplementary Methods. The replications and the prediction market study were approved by the institutional review board or an ethical review board, and participants gave informed consent to participate.

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

Code availability. The analysis codes for both the aggregate data and each individual replication are available at the project's OSF repository (https://osf.io/pfdyw/).

Data availability. The data reported in this paper and in the Supplementary Information are tabulated in Supplementary Tables 3–6. The replication reports (pre-data collection and final versions) and the data and analysis code for each individual replication are available in subprojects organized in the same repository (https://osf.io/pfdyw/).

Received: 6 March 2018; Accepted: 6 July 2018;
Published online: 27 August 2018

References

1. McNutt, M. Reproducibility. *Science* **343**, 229 (2014).
2. Baker, M. Is there a reproducibility crisis? *Nature* **533**, 452–454 (2016).
3. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
4. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
5. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10**, 712 (2011).
6. Begley, C. G. & Ellis, L. M. Drug development: raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
7. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
8. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
9. Maniadis, Z., Tufano, F. & List, J. A. One swallow doesn't make a summer: new evidence on anchoring effects. *Am. Econ. Rev.* **104**, 277–290 (2014).
10. Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. The economics of reproducibility in preclinical research. *PLoS Biol.* **13**, e1002165 (2015).
11. Klein, R. A. et al. Investigating variation in replicability: a 'many labs' replication project. *Soc. Psychol.* **45**, 142–152 (2014).
12. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
13. Camerer, C. F. et al. Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
14. 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).
15. Klein, R. A. et al. Many Labs 2: investigating variation in replicability across sample and setting. *Adv. Methods Prac. Psychol. Sci.* (in the press).
16. Ackerman, J. M., Nocera, C. C. & Bargh, J. A. Incidental haptic sensations influence social judgments and decisions. *Science* **328**, 1712–1715 (2010).
17. Aviezer, H., Trope, Y. & Todorov, A. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* **338**, 1225–1229 (2012).
18. Balafoutas, L. & Sutter, M. Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science* **335**, 579–582 (2012).
19. Derex, M., Beugin, M.-P., Godelle, B. & Raymond, M. Experimental evidence for the influence of group size on cultural complexity. *Nature* **503**, 389–391 (2013).
20. Duncan, K., Sadanand, A. & Davachi, L. Memory's penumbra: episodic memory decisions induce lingering mnemonic biases. *Science* **337**, 485–487 (2012).
21. Gervais, W. M. & Norenzayan, A. Analytic thinking promotes religious disbelief. *Science* **336**, 493–496 (2012).
22. Gneezy, U., Keenan, E. A. & Gneezy, A. Avoiding overhead aversion in charity. *Science* **346**, 632–635 (2014).
23. Hauser, O. P., Rand, D. G., Peysakhovich, A. & Nowak, M. A. Cooperating with the future. *Nature* **511**, 220–223 (2014).
24. Janssen, M. A., Holahan, R., Lee, A. & Ostrom, E. Lab experiments for the study of social-ecological systems. *Science* **328**, 613–617 (2010).
25. Karpicke, J. D. & Blunt, J. R. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science* **331**, 772–775 (2011).
26. Kidd, D. C. & Castano, E. Reading literary fiction improves theory of mind. *Science* **342**, 377–380 (2013).
27. Kovacs, Á. M. & Téglás, E. & Endress, A. D. The social sense: susceptibility to others' beliefs in human infants and adults. *Science* **330**, 1830–1834 (2010).
28. Lee, S. W. S. & Schwarz, N. Washing away postdecisional dissonance. *Science* **328**, 709 (2010).
29. Morewedge, C. K., Huh, Y. E. & Vosgerau, J. Thought for food: imagined consumption reduces actual consumption. *Science* **330**, 1530–1533 (2010).
30. Nishi, A., Shirado, H., Rand, D. G. & Christakis, N. A. Inequality and visibility of wealth in experimental social networks. *Nature* **526**, 426–429 (2015).
31. Pyc, M. A. & Rawson, K. A. Why testing improves memory: mediator effectiveness hypothesis. *Science* **330**, 335 (2010).
32. Ramirez, G. & Beilock, S. L. Writing about testing worries boosts exam performance in the classroom. *Science* **331**, 211–213 (2011).
33. Rand, D. G., Greene, J. D. & Nowak, M. A. Spontaneous giving and calculated greed. *Nature* **489**, 427–430 (2012).
34. Shah, A. K., Mullainathan, S. & Shafir, E. Some consequences of having too little. *Science* **338**, 682–685 (2012).
35. Sparrow, B., Liu, J. & Wegner, D. M. Google effects on memory: cognitive consequences of having information at our fingertips. *Science* **333**, 776–778 (2011).
36. Wilson, T. D. et al. Just think: the challenges of the disengaged mind. *Science* **345**, 75–77 (2014).
37. Bohannon, J. Replication effort provokes praise—and 'bullying' charges. *Science* **344**, 788–789 (2014).
38. Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. Comment on "Estimating the reproducibility of psychological science". *Science* **351**, 1037 (2016).
39. Anderson, C. J. et al. Response to comment on "Estimating the reproducibility of psychological science". *Science* **351**, 1037 (2016).
40. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).
41. 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).
42. Etz, A. & Vandekerckhove, J. A Bayesian perspective on the Reproducibility Project: Psychology. *PLoS One* **11**, e0149794 (2016).
43. Gelman, A. & Stern, H. The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.* **60**, 328–331 (2006).
44. Cumming, G. Replication and *P* intervals: *P* values predict the future only vaguely, but confidence intervals do much better. *Psychol. Sci.* **3**, 286–300 (2008).
45. Verhagen, J. & Wagenmakers, E.-J. Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* **143**, 1457–1475 (2014).
46. Simonsohn, U. Small telescopes: detectability and the evaluation of replication results. *Psychol. Sci.* **26**, 559–569 (2015).
47. 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).
48. Wagenmakers, E.-J. et al. Bayesian inference for psychology. Part II: example applications with JASP. *Psychon. Bull. Rev.* **25**, 58–76 (2017).
49. Lee, M. D. & Wagenmakers, E.-J. *Bayesian Cognitive Modeling: A Practical Course* (Cambridge Univ. Press, Cambridge, 2013).
50. Dreber, A. et al. Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl Acad. Sci. USA* **112**, 15343–15347 (2015).
51. Benjamin, D. et al. Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).
52. Jeffreys, H. *Theory of Probability* (Oxford Univ. Press, Oxford, 1961).

53. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
54. Arrow, K. J. et al. The promise of prediction markets. *Science* **320**, 877–878 (2008).
55. Nosek, B. A., Ebersole, C. R., DeHaven, A. & Mellor, D. M. The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606 (2018).
56. Nosek, B. A. et al. Promoting an open research culture: author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science* **348**, 1422–1425 (2015).

Acknowledgements

Neither *Nature Human Behaviour* nor the publisher had any involvement with the conduct of this study prior to its submission to the journal. For financial support we thank: the Austrian Science Fund FWF (SFB F63, START-grant Y617-G11), the Austrian National Bank (grant OeNB 14953), the Behavioral and Neuroeconomics Discovery Fund (C.F.C.), the Jan Wallander and Tom Hedelius Foundation (P2015-0001:1 and P2013-0156:1), the Knut and Alice Wallenberg Foundation (Wallenberg Academy Fellows grant to A.D.), the Swedish Foundation for Humanities and Social Sciences (NHS14-1719:1), the Netherlands Organisation for Scientific Research (Vici grant 016.Vici.170.083 to E.-J.W.), the Sloan Foundation (G-2015-13929) and the Singapore National Research Foundation's Returning Singaporean Scientists Scheme (grant NRF-RSS2014-001 to T.-H.H.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the following people for assistance with the experiments and analyses: D. van den Bergh, P.-C. Bindra, J. van Doorn, C. Huber, A. Ly, M. Marsman and J. Zambre.

Author contributions

C.F.C., A.D., F.H., J.H., T.-H.H., M.J., M.K., G.N., B.A.N. and T.P. designed the research. C.F.C., A.D., F.H., T.-H.H., J.H., M.J., M.K., D.M., G.N., B.A.N., T.P. and E.-J.W. wrote the paper. T.C., A.D., E.F., F.H., T.-H.H., M.J., T.P. and Y.C. helped to design the prediction market part. F.H. and E.-J.W. analysed the data. A.A., N.B., A.G., E.H., F.H., L.H., T.I., S.I., D.M., J.R. and H.W. carried out the replications (including re-estimating the original estimate with the replication data). All authors approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0399-z.

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

Correspondence and requests for materials should be addressed to B.A.N.

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

nature research

Corresponding author(s): Brian Nosek

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
□	☒	An indication of 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 statistics 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
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	We replicated 21 original studies using the software of the original study whenever possible; in the replications where we used other software this is stated in the SI and the Replication Report for each replication (the Replication Reports and all the softwares used in the replications are available at OSF at https://osf.io/pfdyw/).
Data analysis	We have posted code for all data analyses carried out in the Replication Reports for each replication and for all the analyses in the manuscript and SI at OSF (https://osf.io/pfdyw/).

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

All data for the 21 replications have been posted at OSF (https://osf.io/pfdyw/).

Field-specific reporting

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

□ Life sciences ☒ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences

Study design

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

Study description	Replications of 21 experimental studies in the social sciences (with pre-registration of all the replications).
Research sample	The research samples are similar to the ones used in the original studies (students or other easily accessible adult subject pools in line with our inclusion criteria for studies to replicate) and are described in the Replication Reports for each replication posted at OSF (https://osf.io/pfdyw/).
Sampling strategy	The sampling strategy is similar to the one used in the original studies (students or other easily accessible adult subject pools) and are described in the Replication Reports for each replication posted at OSF (https://osf.io/pfdyw/). For sample sizes we used a two-stage procedure with 90% power to detect 75% of the original effect size at the 5% level (two-sided test) in Stage 1; if the effect was not significant in the original direction in Stage 1 a second data-collection was carried out with 90% power to detect 50% of the original effect size at the 5% level (two-sided test) in the pooled first and second stage data collection.
Data collection	The data collection for all replications was done as similarly as possible to the data collection in the original studies and are described in detail in the Replication Report for each replication posted at OSF (https://osf.io/pfdyw/).
Timing	The data collections for the replications were done between September 2016 and September 2017, and the data collection for the prediction markets were done in November 2016.
Data exclusions	We used the same criteria for data exclusions as in the original studies and any deviations from this are stated in the Replication Reports for each replication posted at OSF (https://osf.io/pfdyw/).
Non-participation	Any non-participation is detailed in the Replication Reports for each replication posted at OSF (https://osf.io/pfdyw/).
Randomization	The experimental procedures, including the randomization, follow the original studies and are detailed in the Replication Reports for each replication posted at OSF (https://osf.io/pfdyw/).