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 $BF_{+0}=6.8$, 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.

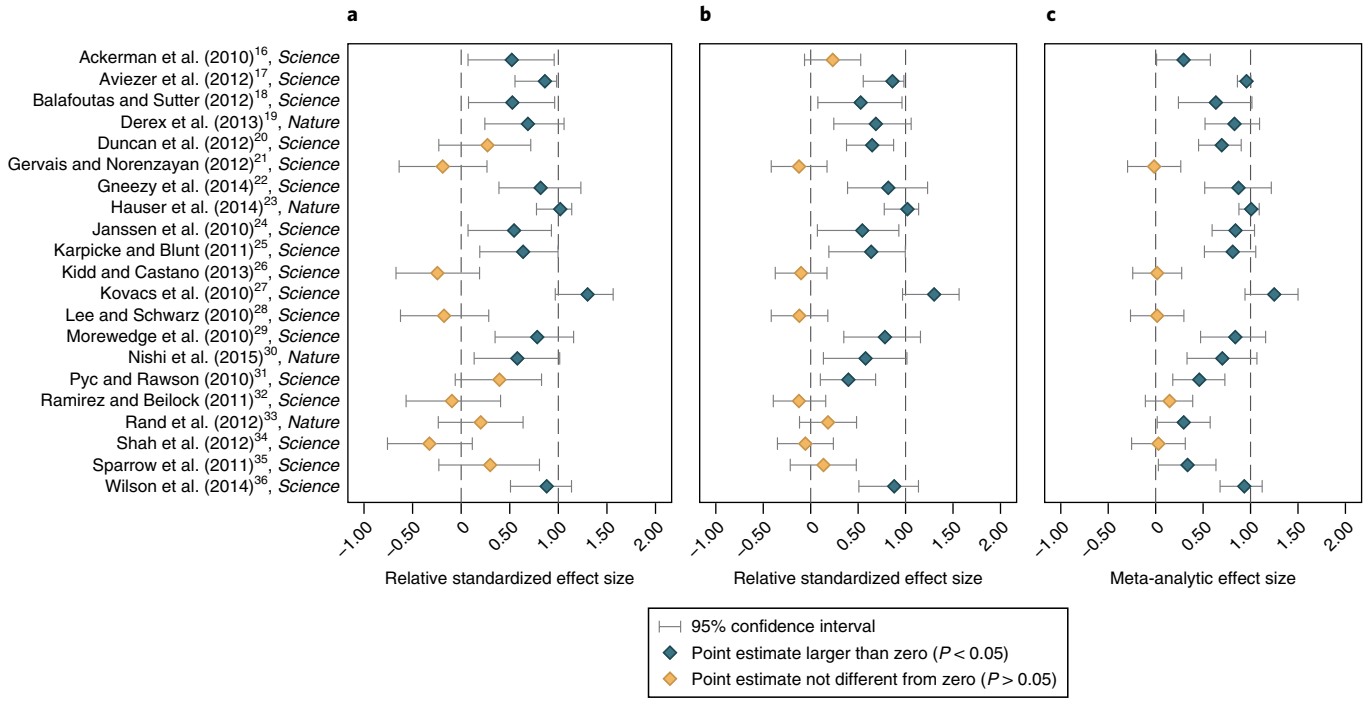



**Fig. 1 | Replication results after stage 1 and stage 2. a**, Plotted are the 95% CIs of the replication effect sizes (standardized to the correlation coefficients *r*) after stage 1. The standardized effect sizes are normalized so that 1 equals the original effect size. There is a significant effect 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.

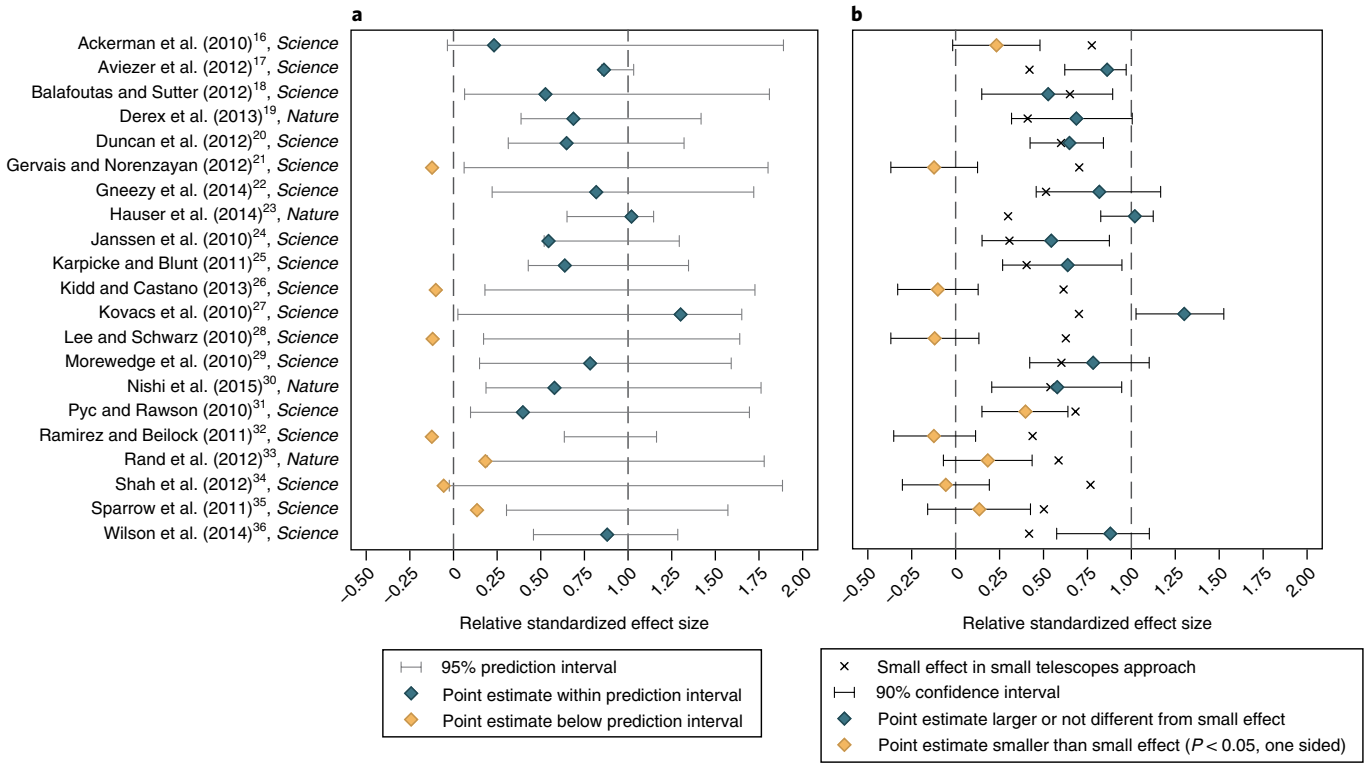



**Fig. 2 | Replication results for two complementary replication indicators. a**, Plotted are the 95% prediction intervals[47] for the standardized original effect sizes (correlation coefficient *r*). The standardized effect sizes are normalized so that 1 equals the original effect size. Original and 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

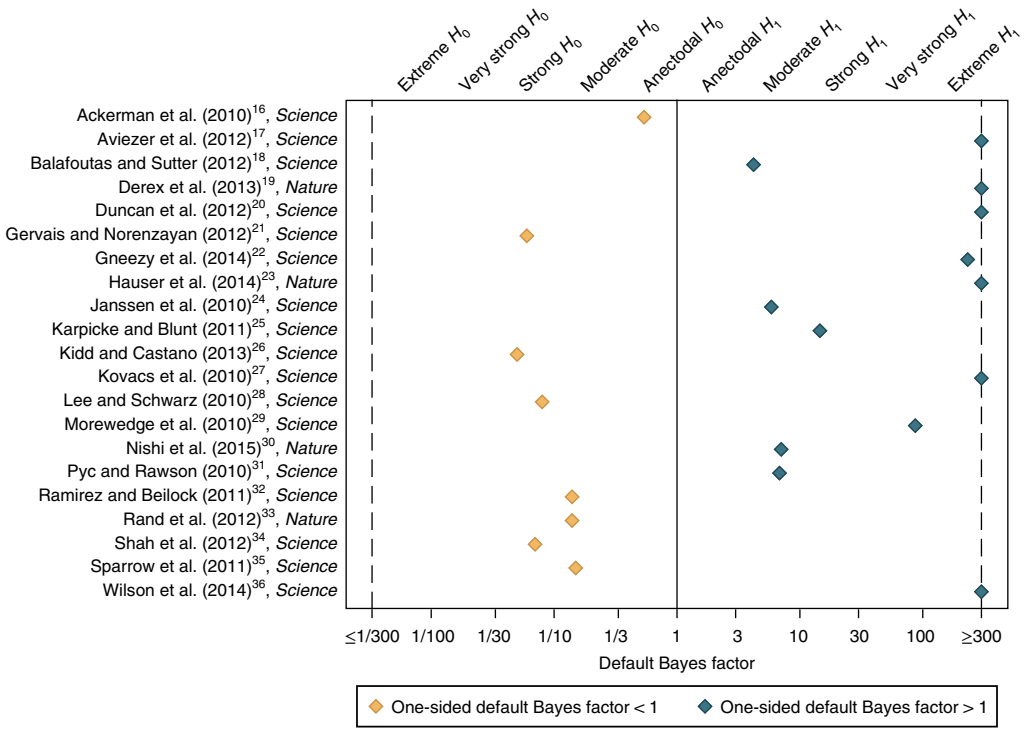



**Fig. 3 | Default Bayes factors (one sided) for the 21 replications.** A default Bayes factor[48] above 1 favours the hypothesis of an effect in the direction of the original paper and a default Bayes factor below 1 favours the null hypothesis ($H_0$) of no effect. The evidence categories proposed by 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,

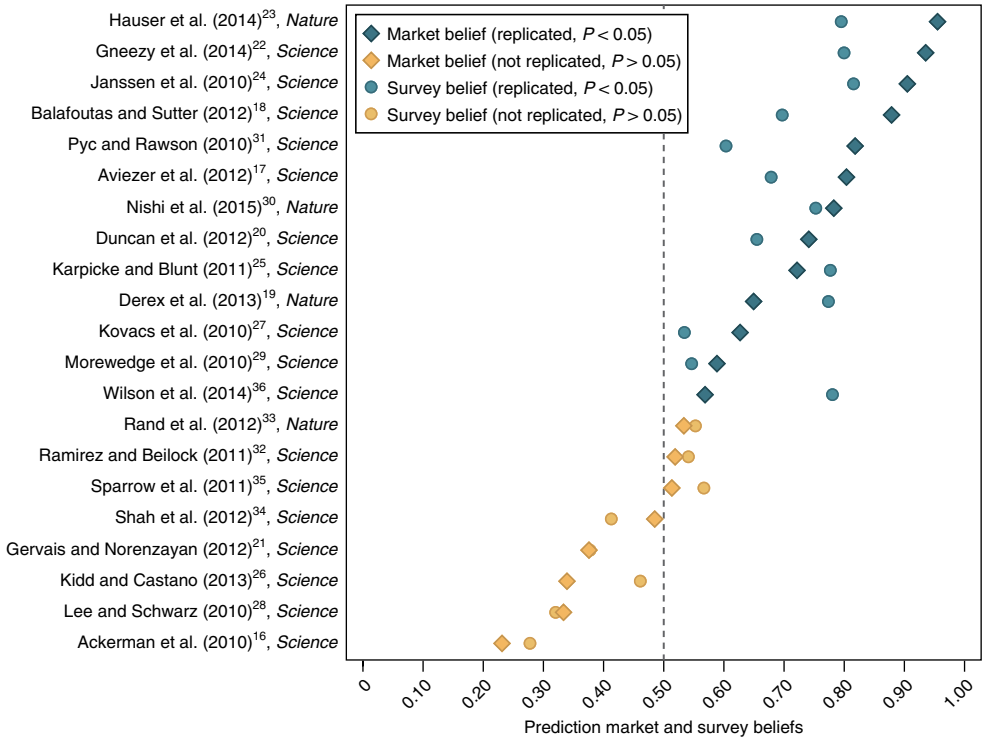



**Fig. 4 | Prediction market and survey beliefs.** The prediction market beliefs and the survey beliefs of replicating (from treatment 2 for measuring beliefs; see the Supplementary Methods for details and Supplementary Fig. 6 for the results from treatment 1) are shown. The replication studies are ranked 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 human behaviour

SUPPLEMENTARY INFORMATION

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

In the format provided by the authors and unedited.

# 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]**

[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

Supplementary Information for

# Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer[†], Anna Dreber[†], Felix Holzmeister[†], Teck-Hua Ho[†], Jürgen Huber[†], Magnus Johannesson[†], Michael Kirchler[†], Gideon Nave[†], Brian Nosek[†,*], Thomas Pfeiffer[†], Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, Hang Wu

[*] To whom correspondence should be addressed. E-mail: nosek@cos.io.
[†] These first ten authors contributed equally to this work.

**The Supplementary Information contain:**

# Supplementary Methods

Here we provide further details on the replications, the estimation of standardized effect sizes and complementary replicability indicators, the implementation of the prediction markets and surveys, the comparison of prediction market beliefs, survey beliefs, and replication outcomes, the comparison of reproducibility indicators to experimental economics and the psychological sciences, and additional results and data for the individual studies and markets. The code used for the estimation of replication power, standardized effect sizes, all complementary replication indicators, and all results is posted at OSF (https://osf.io/pfdyw/).

## Replications

### Inclusion criteria

We replicated 21 experimental studies in the social sciences published between 2010 and 2015 in *Nature* and *Science*. We included all studies that fulfilled our inclusion criteria for: (i) the journal and time period, (ii) the type of experiment, (iii) the subjects included in the experiment, (iv) the equipment and materials needed to implement the experiment, and (v) the results reported in the experiment. We did not exclude studies that had already been subject to a replication, as this could affect the representativity of the included studies. We define and discuss the five inclusion criteria below.

**Journal and time period:** We included experimental studies published in *Nature* and *Science* between 2010 and 2015. The reason for focusing on these two journals is that they are typically considered the two most prestigious general science journals. Articles published in these journals are considered exciting, innovative, and important, which is also reflected in their high impact factors.

**Type of Experiment:** We included experimental studies using within or between subjects treatment comparisons to estimate a treatment effect. Between and within subjects treatment effects are both designs to identify causal effects. The between subjects treatment design is the "classical" design used in randomized controlled trials (RCT), where participants are randomly allocated to two or more treatments and the outcome is compared between treatments. This design is for instance the gold standard in medicine in comparing different medical treatments. Another commonly used design is a within subjects treatment comparison where the same participants are exposed to two or more treatments and the outcome is compared between the treatments. We did not include studies that compare behaviors across different groups of participants (e.g., men vs. women or high vs. low income individuals). Such studies describe how behavior differs across groups, but they do not allow for causal interpretation (unless strong assumptions are made). We also did not include studies comparing the behavior of a group of individuals with a theoretical prediction. Such studies are also commonly referred to as "experiments", but do not estimate any treatment effects.

**Subjects included in the experiment:** We included experiments using students or other easily accessible adult subject pools. These include online experiments on convenience samples such as experiments using Amazon Mechanical Turk (AMT). This inclusion criterion implies that field experiments are not included, and the reason for this is that field experiments are typically substantially more expensive and difficult to conduct. Our inclusion criteria also imply excluding experiments on children.

**Equipment and materials needed to implement the experiment:** We included experiments that can be conducted in a standard lab used for experiments in the social sciences (an experimental lab in psychology or economics). This implies that we exclude experiments requiring non-standard equipment, interventions, or materials. This means that for instance fMRI studies are excluded. Examples of other studies excluded not meeting this inclusion criteria are the study by Falk and Szech[1] using mice and the study by Gelstein et al.[2] using human tears. This inclusion criteria is motivated by feasibility and funding constraints.

**Results reported in the experiment:** We included experiments that test at least one clear hypothesis with a statistically significant finding ($p < 0.05$). This implies that we do not include studies reporting null findings, but we focus on replicating findings reported as "positive findings".

All the papers meeting these inclusion criteria were included ($n = 21$). We did not exclude studies that had already been subject to a replication, as this could affect the representativity of the included studies (in one case discussed further below, an ongoing Registered Replication Report study, along with other reasons, affected which study within that paper was selected for replication). Many of the included papers reported more than one tested hypothesis with a significant finding, and more than one study. We therefore used the following three criteria in descending order to determine which treatment effect to replicate.

1. Include the first study reporting a significant treatment effect for papers reporting more than one study.

2. Include the most central and important statistically significant result (as emphasized in the original study) of this study among all within or between subjects treatment comparisons. The interpretation of which was the most central and important result in the selected study was made by us, but the original authors could provide feedback on this when they commented on the draft of the replication report sent to them for feedback (see below). Prior to conducting the replications we did not receive feedback from any of the original authors, that we had selected the "wrong" result. After conducting the replication and seeing the results the original authors in one study argued that we had not selected the most central and important result. This was for the replication of Janssen et al.[3], where the original authors ex post argued that comparing the effect of communication on net earning within subjects was a more central result of their study (in our replication we replicated the result on the effect of communication on net earnings between subjects as the between subjects comparison is a stronger identification of causal effects). In some other cases the original authors of papers reporting the results of a series of studies argued that the first study selected for replication did not contain the most important result in their

paper; but in these cases the result selected for replication followed directly from applying criteria 1 above.

3. If more than one equally central result remained, we randomly picked one of the results for replication. This was the case for three papers[4–6].

The first criterion differs from the criteria used in the Replication Project: Psychology (*RPP*)[7] and the Experimental Economics Replication Project (*EERP*)[8]. In those projects the last study, rather than the first study, was included for studies reporting a series of studies. To include the last study (or the first study) is an arbitrary rule for selecting which study to replicate, and there is little a priori reason to believe that replicability differs depending on the order studies are reported. However, it may be problematic if a rule is established that replications are always conducted on the last study, as that may lead researchers to present the study they believe the most likely to replicate, last in the paper. To counteract such possible strategic incentives we decided to include the first rather than the last study for papers reporting a series of studies.

We deviated from that criteria for one study: the Rand et al.[9] paper. The Rand et al.[9] paper included 10 studies. Studies 1–5 were correlational studies not estimating treatment effects, and studies 6–7 tested the effect of time pressure on contributions to a public good; study 7 is included in a recently published Registered Replication Report study[10] (study 6 is similar to study 7, but conducted on AMT instead of in the lab). Based on criteria 1 above, study 6 would have been selected for replication as it was the first treatment effect study, but as this study was already subject to a Registered Replication Report study organized by *Perspectives of Psychological Science*, we chose not to select this study for replication (one of the authors of this project was also involved in one of the replications of the Registered Replication Report). A further reason for not replicating study 6 or 7 in Rand et al.[9] is that it has been noted that the reported significant finding was due to selection bias and analyzing the data appropriately as "intention to treat" did not yield a significant effect[11]. The intention to treat result was also the primary focus of the Registered Replication Report[10]. The Registered Replication Report found that the result did not replicate, but this was not known to us when we decided not to replicate this study. For the above reasons we decided to include study 8 in Rand et al.[9], as it was

the first treatment effect study not subject to an ongoing Registered Replication Report study and it was not subject to a controversy over the appropriate analysis of the data. Note also that for some of the other papers our criteria above implied that study 1 was not replicated, as it did not report any significant treatment effects. This was the case for Duncan et al.[12] in which we included study 1b, Gervais and Norenzayan[13] in which we included study 2, and Wilson et al.[14] in which we included study 8.

In our initial screening we selected 22 studies that met the above inclusion criteria. We contacted the original authors requesting materials and asking for feedback on an initial version of the replication report for these 22 studies. However, during this process we discovered that the result selected for replication for one of these studies did not fulfill our inclusion criteria and this study was excluded. This was the study by Mani et al.[15]. Originally we intended to replicate the result that inducing thoughts about financial hardship reduces cognitive function more among poor individuals than rich ones (from study 1 in the paper). This was an interaction effect between a treatment that induced thinking about a hard versus an easy financial problem, and individual differences in participants' income (i.e., above or below median household income). But as the income level was not assigned to households as a randomized treatment, but rather was based on their actual income, this does not match the definition of a treatment effect; it is instead a comparison of a treatment effect between different groups (rich and poor participants) and the comparison cannot be given a causal interpretation (unless strong assumptions are made). It is thus not consistent with our criteria for the type of experiment (a between or within subjects treatment effect) which was only realized after the initial screening. It was also a borderline case in terms of the subjects included in the experiment criteria (the experiment was conducted at a mall with subjects recruited from the mall). This study was thus excluded from the replication project prior to starting any data collection and our final replication sample consisted of 21 studies. If an original study included more between subjects treatments than the treatments selected for replication, we in general only included the treatments used for the result selected for replication.

## Replication procedure

**Statistical Tests.** In the replications, we used the same statistical test as was used in the original study. For one study[16] we deviated from this rule as it turned out that the exact same test as used in the original study yielded clearly implausible standard errors on the replication data, stemming from the two-way clustering (on the individual and round level) of standard errors used in the original paper; we instead modified the test slightly and used one-way clustering (on the individual level) as the main replication test. This, however, does not affect the conclusion about whether the study replicates or not. See the replication report for this study for further details. As we used the same statistical tests as used in the original studies, we did not test if the data collected in the replications met the assumptions of the statistical tests used (such as assumptions about normality or equal variances). Thus, for tests based on normal distributions the data distribution was assumed to be normal but this was not formally tested. All reported $p$-values in the paper and Supplementary Information are based on two-sided tests.

**Replication teams.** There were five replication teams: a team at CalTech and Wharton (responsible for four replications); a team at the Center for Open Science and the University of Virginia (responsible for five replications); a team at the Stockholm School of Economics (responsible for five replications); a team at the National University of Singapore (responsible for three replications); and a team at the University of Innsbruck (responsible for four replications). All the online experiments using AMT were replicated by the same team (the team at the Stockholm School of Economics).

**Language.** One of the original experiments was conducted in German, one in French and one in Italian; the remaining ones were conducted in English. The replication of the original study in German was conducted in German (the team at the University of Innsbruck), but the replications of the original studies in French and Italian were conducted in English. Fifteen out of the 18 original experiments in English were replicated in English and the remaining three studies were replicated in German (the team at the University of Innsbruck).

**Experimental software.** The same software and computer programs as in the original experiments were used to conduct the replications whenever possible, but for some studies we needed to program new software to conduct the replications:

- The experiment in Aviezer et al.[17] was implemented using the software package *E-Prime*. While the original authors were eager to provide us with the original code, the replication experiments were implemented using a custom-designed software module using *oTree*[18]. However, instructions used in the original experiment were provided and the final program was approved by the original authors.
- The original authors were unable to retrieve the *E-Prime* computer program used in the original experiment by Karpicke & Blunt[19]. Specifically, in the original study, the sea otters text was presented on the computer screen, as was the answer box for the retrieval practice condition. However, all other tasks were completed using pencil and paper in the original study. For this reason, our replication used hard copies of all of the original materials used in the original study (including the sea otters text and the retrieval practice answer box originally run through the *E-Prime* program). The original authors agreed that the change from computer to paper responses should not be a consequential difference for the effect studied in the replication.
- In Kovacs et al.[20] the original experiment was carried out in *Psyscope X* to measure reaction times, but the original program is no longer available. Therefore we developed a computer program to measure reaction times used in the replication.
- The software used in the original experiment of the study by Morewedge et al.[6], *Adobe Authorware*, is no longer supported. The replication experiment was implemented using *z-Tree*[21]. However, instructions and pictures used in the original study were made available by the original authors and the final software used in the replication experiment was approved by the original authors.
- The original study by Nishi et al.[16] was conducted using a software called *Breadboard*, but the version used by the original authors is no longer supported (and the original authors did not provide the code for the version used in the original experiment). We therefore programmed the experiment from the beginning in the most recent version of *Breadboard*.

- The original authors of Ramirez & Beilock[22] did not provide the software used in the original study, and the experiment was therefore programmed in *Qualtrics* based on the materials and instructions provided by the original authors.

- For the replication of Rand et al.[9], *Qualtrics* was used instead of *LimeSurvey* using the same phrasing as in the original experiment (and this was approved by the original authors).

- For the replication of Sparrow et al.[23], the original authors did not respond to our requests for feedback and the original materials and software used, and the senior author passed away in 2013. We therefore programmed the software for this replication ourselves using *oTree*[18] based on the information provided in the original paper and the supplementary materials. But as we received no feedback about the design from the original authors, we could not verify that the instructions and implementation of the experiment in the replication were identical to the original study. After the replication had been conducted we received some feedback from the original authors detailing the following design differences. (i) The cognitive load manipulation differed between the original study and the replication, such that in the original study participants were asked to remember a different six-digit number for each word used in the modified Stroop task, whereas in the replication participants were asked to remember the same six-digit number while completing all trials of the modified Stroop task. (ii) The replication used 48 trials of the modified Stroop task whereas the original study used 24 trials. (iii) The replication excluded mistakes from the analysis (where participants click on the key for the wrong color), whereas this was not done in the original study. However, the conclusion about the study not replicating remains unchanged when only the first 24 trials are included in the analysis or if mistakes are not excluded from the analysis. It also turned out that the number of participants reported in the original paper was not correct, which affected the power of this replication to some degree (see below). While we consider it implausible, we cannot rule out the possibility that the difference in the cognitive load manipulation impacted the replication result.

**Replication reports and pre-registration.** The replication team responsible for each replication wrote a replication report detailing the planned replication (with the following sections: hypothesis to replicate and bet on; power analysis and criteria for replication: first data collection; power analysis and criteria for replication: second data collection; sample; materials; procedure; analysis; and differences from the original study). A draft of each replication report was sent to the original authors for feedback, and the replication reports were revised based on the comments; this process continued until the original authors approved the replication report (except for the Sparrow et al.[23] replication, where the original authors did not provide any feedback on the replication report and the Ramirez & Beilock[22] replication where the original authors after seeing the replication results argued that they had not approved the replication report). This version of each replication report (the pre-replication version) was then posted at the project webpage (www.socialsciencesreplicationproject.com) and pre-registered at the Open Science Framework (OSF, https://osf.io/pfdyw/) prior to starting the data collection (we also saved all communications between the original authors and the replication teams on a dedicated e-mail account). After the replications had been conducted, the replication reports were updated with the results of the replication (the following four sections were added to the reports: results for the first data collection (90% power to detect 75% of the original effect size); results for the first and second data collection pooled (90% power to detect 50% of the original effect size); unplanned protocol deviations; and discussion. The replication reports with the updated results were again sent to the original authors for comments. After revision, the final versions of the replication reports were posted at www.socialsciencesreplicationproject.com and at OSF (https://osf.io/pfdyw/) (the versions prior to the replications and the final versions are posted and publicly available). In Supplementary Table 2, which lists each study, we document whether the original authors of that study shared the materials of the original study and whether or not they approved of the pre-replication version of the Replication Report with the pre-registered design of the replication. In the table we also note if the replication used the same software as the original study. We used the same exclusion criteria as used in the original studies for excluding any observations in the analyses; any deviations from this are detailed in the "Differences from the original study" section (if the deviation was planned

and pre-registered) or the "Unplanned protocol deviations" section (if the deviation was unforeseen and made after the data collection) in the replication reports. At the end of the Supplementary Methods we list any "unplanned protocol deviations" for each of the replications.

We invited all original authors to post a comment about the replication alongside the replication reports at the project website and OSF. For several studies comments by the original authors have been posted at the time of writing this SI. In these comments the original authors have noted some perceived limitations on the replications of their studies. Also, we received some comments in the journal peer review process. These are briefly discussed below (the comments of the original authors of Sparrow et al.[23] have also already been discussed above).

The Duncan et al.[12] study replicated in Stage 2 based on the statistical significance criterion, but with a smaller effect size than observed in the original study. Here the original authors observe that the memory accuracy was lower in the replication than in the original study, and argues that this implies higher measurement error in the replication and that this may have lowered the replication effect size. However, the lower memory accuracy in the replication may be due to a lower memory ability in the replication sample, and we cannot differentiate between these two explanations in the data. The original authors also note that the replication results show larger treatment effects when analyzed according to the subjective mnemonic reports (i.e. based on the subjective evaluation of whether the preceding trial was new or old instead of the objective classification). Based on the subjective mnemonic reports the result is already significant in Stage 1 ($p = 0.035$), and the results after Stage 2 are more significant than in the main replication test based on the objective classification ($t(91) = 5.78$ with the subjective classification and $t(91) = 4.63$ with the objective classification). The results based on the subjective classification are reported in the replication report for the study and are consistent with the conclusion of the original study replicating based on the pre-registered test for the objective classification.

Janssen et al.[3] whose study replicated according to the statistical significance criterion, but with a smaller effect size than in the original study, argues that the result tested is not the most important statistically significant finding of their study (the result tested is

that communication increase average earnings in a common-pool resources game). This feedback was not communicated to us prior to conducting the replication when we made a priori commitments to identifying the key finding. The original authors argue that the effect of communication was of main interest to their study, but that the within-subject test of the effect of communication was more central to their study than the between-subjects test of communication we replicated. We originally decided to replicate the between-subjects test rather than the within-subjects test, as the identification of the causal effect is stronger in the between subjects test.

The original authors of Kidd & Castano[24] acknowledge that our replication is a direct replication of their study 1, with the methods closely following those of the original study. However, they argue that the design used in their study 1 has important limitations, and that subsequent studies have improved on the methodology through better methods of removing individuals who did not read their assigned texts (methods of excluding individuals with low reading times). They therefore argue that our replication is a relatively poor test of the underlying hypothesis. Here we note that we followed their original procedures closely, and the replication results fail to support the original findings and provide strong support for the null hypothesis. We are also not convinced that removing observations from the data is an improvement in methodology, but that intention to treat is in general a preferred methodology for identifying causal effects (especially in this case as the reading time may be endogenous to the treatment, causing selection bias if individuals with low reading times are excluded).

Lee & Schwarz[25] argue that an unintended methodological difference between the original study and the replication may have affected the results of the replication. In the original experiment, participants examined covers of 30 music albums, picked 10 they would like to own, wrote down these 10 albums' titles and artist names, and reported their pre-manipulation rankings for these 10 albums in order of preference. Next, participants worked on filler tasks while the experimenter prepared a different form by listing the 10 albums in alphabetical order of the artist names (rather than album titles). The presentation order of albums did therefore differ in the pre- vs. post-manipulation evaluation and the original authors argue that this is important for the design as they argue that easy recall of one's pre-manipulation evaluation may constrain one's post-manipulation

evaluation. In the replication, we did by mistake not follow this step during the data collection. Instead, the presentation order of albums in the post-manipulation evaluation was the same as in the pre-manipulation evaluation if participants copied them in that order. The original authors argue that this design difference compared to the original study may have reduced the chances of finding a significant effect of washing away postdecisional dissonance. The original authors argue that postdecisional dissonance is a phenomenon that involves multiple processes, and that the unintended methodological deviation in the replication may have increased reliance on one's memory of the earlier evaluation at the expense of the affective processes that motivated the work of the original authors on washing away postdecisional dissonance. We do not dismiss the possibility that the protocol deviation could have had an impact on the observed effects of dissonance reduction, but we may find that possibility less plausible than do the original authors. We did replicate the dissonance effect, and we did not observe any reduction of that dissonance effect whether or not participants copied in the order of ranks or not (which we perceive as a plausible reinforcement of memory for the original ranks). Nevertheless, we cannot conclusively determine the reason for this, or any of the other, failures to replicate in the *SSRP*. As the original authors suggest follow-up investigation is necessary to firmly determine if the unintended protocol deviation in the replication is important or not.

The original authors of Ramirez & Beilock[23] argue that we in the replication failed to create a high-pressure performance situation because there was no significant difference in performance between the high pressure and the low pressure condition and that we could therefore not test whether expressive writing can boost performance under pressure. We disagree with the interpretation of the original authors. As suggested by the original authors we preregistered tests of manipulation checks to assess whether the pressure manipulation increased participant anxiety, and those manipulation checks were successful. For further discussion on this we refer to the original author comment on this replication and the email correspondence between the original authors and the replication team. These appear at https://osf.io/hps2b/ and https://osf.io/n276s/ respectively.

The original authors of Rand et al.[9] raise the issue that the Amazon Mechanical Turk (AMT) subject pool has changed since the data collection was carried out in their study, and in particular that the experience of economic game paradigms have increased over

time in the AMT subject pool. Prior to conducting the replication they also asked us to include a question about prior experience in the data collection, which we did. But they did not ask us to pre-register any analyses or robustness test using this question. At their advice prior experience was measured on a 5-point likert scale; with 1 (nothing similar), 3 (something similar) and 5 (exactly this scenario) marked on the scale. The distribution of the $n = 2,136$ responses on this scale was: 1 ($n = 367$), 2 ($n = 362$), 3 ($n = 1,033$), 4 ($n = 282$), and 5 ($n = 92$); with a mean of 2.71. The experience level can not directly be compared to the replicated study in Rand et al.[9] as experience was not measured in that study, and without pre-registering how to define low and high experience it is not obvious what cut-off to use on the five-point scale. After seeing the data, the original authors did a sub-group analysis on those with the lowest experience level (1 on the scale). They found a similar effect size in this sub-group as for the overall sample in the original study, but it was non-significant ($p = 0.108$) and therefore provides no evidence for an effect among inexperienced subjects (the $n = 367$ limits the power of this test, although it is higher than the $n = 343$ in the original study; as the test was not pre-registered it should furthermore be interpreted cautiously). It cannot be ruled out that changes in the AMT subject pool over time affects results, but we also note that the two other studies on the public goods game based on AMT data replicated (Hauser et al.[26] and Nishi et al.[16]).

Shah et al.[27] carried out a series of five studies and we replicated study 1 based on our criteria for replicating the first study reporting a significant treatment effect. In their comments the original authors argue that study 1 is less important for their paper than their other studies, and inspired by our replication they decided to carry out a replication study of their own on all their five studies (with results posted at https:osf.io/vzm23/). They did replicate what they consider to be their most important finding in their paper, that scarcity itself leads to over-borrowing. But importantly they also failed to replicate study 1 confirming our findings for that study.

**Separation between prediction markets and replications.** None of the researchers involved in carrying out the replications received any information about the prediction markets results or the survey results until all replications had been conducted. Only two members of the research team (Thomas Pfeiffer and programmer Taizan Chan) had access to information about the prediction market results prior to the completion of the repli-

cations. These two people were not involved in any replication data collection. Everyone involved in carrying out the replications was also instructed not to discuss the prediction market with any individuals who participated in the prediction markets and surveys. This was done to rule out the possibility that the persons conducting the experiments and carrying out the replications were affected by the prediction market results.

**Determination of replication sample sizes.** We used a two-stage procedure for carrying out the replications. In the first data collection 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 the first data collection (a two-sided $p$-value $< 0.05$ and an effect in the same direction as the original study), no further data collection was carried out. If the original result did not replicate in the first data collection, we carried out a second data collection to have 90% power to detect 50% of the original effect size for the first and second data collection pooled. We then tested if the original result replicated in the pooled sample (a two-sided $p$-value $< 0.05$ and an effect in the same direction as the original study).

For the Sparrow et al.[23] replication, the replication power was lower than the planned 90% power (82.0% in the first data collection and 80.7% in the first and second data collection pooled). This is due to an error in the reported sample size in the Sparrow et al.[23] paper; the paper reports a sample size of $n = 46$ but a sample size of $n = 69$ was actually used (the sample size affects the power estimation as it affects the estimated standardized effect size ($r$) of the original study). The original authors only provided feedback about this error in the original paper after the replication had already been conducted.

Due to an initial mistake in analyzing the results of the first data collection in one study[28], a second data collection was carried out in spite of the fact that the first data collection showed an effect in the same direction as the original result (and a two-sided $p$-value $< 0.05$). An initial incorrect analysis of the data from the first data collection showed a $p$-value $> 0.05$, and when this error was detected the second data collection was almost completed (and we therefore decided to complete the second data collection). As the result with all data collected is the most informative, we include the second data collection in the Stage 2 results for this study, although, according to the initial protocol, the second data collection should not have been carried out.

Note that if a study fails to replicate in the first data collection, it is given a second chance to replicate and two tests are conducted. This increases the false positive risk somewhat compared to carrying out a single test. However, as the replication tests are directional (i.e., the effect in the replication needs to be in the same direction as the original study) and two-sided tests are used, the false positive risk in each test is only 2.5% (so the total false positive risk with our two-stage procedure does still not exceed 5%; we ran a simulation to estimate the false positive risk more exactly with our two-stage procedure and the false positive risk is 4.2%). Related to this our power estimations, of the power to detect 50% of the original effect size in the first and second data collection pooled, is somewhat conservative as it does not take into account the dependency of the Stage 1 and Stage 2 tests. We ran a simulation to estimate the power more exactly of our two-stage testing procedure and the power to detect 50% of the original effect size is 91.1% instead of 90%. In some replications the statistical power was slightly larger than 90% as the total number of observations needed to be evenly divisible by some number (e.g., subjects or groups needed to be evenly divided into treatments). In some replications the sample sizes were also slightly larger than the planned sample size (as some original studies used exclusion criteria and the number of exclusions were not known in advance, it was difficult to collect exactly the planned number of observations).

The replication sample size needed was estimated in the same way for all the replications. The standardized effect size was estimated as the correlation coefficient ($r$) for all the original studies in the same way as done for the $RPP$[7] and the $EERP$[8]; see below for more details on this. For the first stage data collection we estimated the sample size needed to have 90% power to detect 75% of the original effect size (expressed as the correlation coefficient $r$) at the 5% level in a two-sided test; in the second stage data collection we estimated the sample size needed to have 90% power in the pooled first and second data collection to detect 50% of the original effect size (expressed as the correlation coefficient $r$) at the 5% level in a two-sided test. Note that using some other measure of effect size such as Cohen's $D$ can result in somewhat different sample sizes as the correlation coefficient ($r$) and Cohen's $D$ are not linearly related.

The replication power and sample sizes are substantially more ambitious than for the *RPP*[7] and the *EERP*[8]. Those projects were based on having 90% power to detect 100% of the original effect size, compared to having 90% power to detect 75% (the first data collection) and 50% of the original effect size in this project (for the first and second data collection pooled). Basing the power calculations on 75% of the original effect size rather than 100% approximately doubles the replication sample sizes, and basing the power calculations on 50% of the original effect size rather than 100% leads to an approximate fourfold increase in sample sizes.

The reason for increasing the power is that the *RPP*[7] and the *EERP*[8] have been criticized for being underpowered as they do not take into account that original effect sizes of true positive findings may be inflated due to publication and reporting biases. We therefore took into account the possibility for inflated effect sizes in original studies with true positive results in our power estimation. The reason for basing the power estimations on having 90% power to detect 50% of the original effect size was based on the replication effect sizes in the *RPP* being about 50% of the original effect sizes on average[7]. Note, however, that this gives an overestimation of the inflation effect as the true average inflation rate should be estimated only for the original studies with true positive findings (but this fraction is not exactly known). If, for instance, the rate of false positives in the *RPP* sample would be 50% (such that the true effect size is zero for 50% of the studies), the *RPP* result would imply zero inflation on average in original studies of true findings. As part of the results of the *SSRP* we estimate the inflation rate in original effect sizes of true positive findings. The estimated relative effect size of the 13 studies that replicated according to the statistical significance criterion is 0.74 with a 95% confidence interval between 0.60 and 0.89. The corresponding estimate from the Bayesian mixture model is 0.71 with a 95% credible interval that ranges from 0.58 to 0.83. These results suggest that we are well powered to detect true positive findings in the *SSRP* (in Stage 2 we have 90% power to detect effect sizes below the 95% confidence intervals in the above estimations).

## Estimation of standardized effect sizes and complementary replicability indicators

**Relative effect size.** Both the *RPP*[7] and the *EERP*[8] used the correlation coefficient ($r$) as a standardized effect size measure to compare effect sizes between original studies and replications. We used the same measure of standardized effect sizes and transformed effect sizes into correlation coefficients ($r$) in the same way as done for the *RPP*[7]. The standard errors of the correlation coefficients were calculated by applying the Fisher transformation, and depend only on the sample size of the study. The correlation coefficient was coded as positive for the original study regardless of the actual sign of the effect, and the replication effect size was coded as positive if it was in the same direction as in the original study, and negative if it was in the opposite direction. In Supplementary Figure 4, we show the relationship between the original and replication standardized effect sizes ($r$).

The standardized effect sizes ($r$) are useful for comparing results between the original and the replication study and to estimate a measure of the relative effect size of the replication. However, it is less useful for comparing the level of effect sizes across studies if the studies are based on different levels of aggregation. Some studies use observations at the group level as the unit of observation and any aggregation reduces the variance of the data (i.e., the variance between individuals is larger than the variance between aggregated groups). A higher degree of aggregation of the data and thus lower variance generally increases the standardized effect size. This is, however, not a problem for comparing the standardized effect size between the original study and the replication, because we carry out the statistical test in an identical way for both the original study and the replication with the same level of aggregation. Due to the difficulties in comparing effect sizes between original studies, we normalized the original effect sizes to 1 in our figures (i.e., the standardized replication effect sizes ($r$) as well as the upper and lower bound of the 95% CI of $r$ are divided by the standardized effect size ($r$) in the original study). However, in Supplementary Table 3 and Supplementary Table 4 we also report the standardized effect sizes ($r$) of the original studies and the replications without the normalization of the original effect size to 1.

**Meta-analytic effect size.** We also computed a fixed-effect weighted meta-analytic effect size measure for each study pair as it was done both for the *RPP*[7] and the *EERP*[8]. The meta-analysis treats original and replicated studies equally except for the difference in sample size and gives an estimate of the pooled effect of the original study and the replication. The meta-analytic effect size measure is based on the assumption that there are no publication or reporting biases in the original studies and should thus be interpreted with great caution. The meta-analytic results are shown in Fig. 1c.

**Prediction intervals.** Patil, Peng and Leek[29] recently proposed another method to assess replicability. They suggest to estimate a 95% prediction interval for the original estimate and test how many of the replications fall within this prediction interval. The method takes into account the variability in both the original study and the replication study. For original studies with relatively high, but "significant", $p$-values prediction intervals can be expected to lead to higher replication rates than our primary replication indicator (a significant effect in the original direction). This is because for original studies with a $p$-value close to 0.05 the prediction interval will overlap with the zero effect size or be close to a zero effect size even for large replication sample sizes (the prediction market formula involves taking the square root of the sum of the variance of the estimated effect size ($r$) in the original study and the replication study; only including the variance of the original estimate of $r$ leads to a lower bound of zero for the prediction interval if the $p$-value of the original study is 0.05; adding also the variance of the replication result widen the prediction interval further). For a false positive with an original $p$-value close to 0.05 the likelihood of the point estimate of the replication falling within the prediction interval is thus high as a false positive has a 50% probability of a point estimate above zero. In other words, with $p$-values close to 0.05 in original research, the prediction interval approach is a very liberal strategy for estimating replication success. However, there is also an effect in the opposite direction for high powered replication designs. For original studies that are highly statistically significant it is possible that a replication can show a significant effect in the same direction as the original study, but that fall below the prediction interval (and is thus classified as not replicating according to the prediction interval replication indicator). The 95% prediction interval results are shown in Fig. 2a.

**"Small Telescopes" approach.** The estimated standardized effect sizes were also used to estimate replicability using the "Small Telescopes" approach, which was recently proposed[30]. In the Small Telescopes approach it is estimated whether the replication effect size is significantly smaller (with a one-sided test at the 5% level) than a "small effect" in the original study. A small effect is defined as the effect size the original study would have had 33% power to detect. If the effect size in the replication is significantly smaller than this "small effect size" it is considered a failed replication (and otherwise it is considered a successful replication). The Small Telescopes approach recommends using a replication sample that is always 2.5 times the original sample size, as this gives about 80% power to reject a "small effect"[30]. On average, the replication sample sizes in the *SSRP* are larger than the sample sizes proposed in the Small Telescopes approach, leading to higher power on average to reject a "small effect". The replication sample sizes in Stage 1 are on average about 3 times as large as the original sample sizes and replication sample sizes in Stage 2 are on average about 6 times as large as the original sample sizes. With high powered replication designs the results with the Small Telescopes approach and our primary replication indicator can be expected to start converging. The Small Telescopes indicator can even lead to a lower replication rate as with a high powered replication it is possible to find a significant effect in the same direction as the original study, but that is significantly smaller than a "small effect size". The results for the Small Telescopes approach are shown in Fig. 2b.

**Bayesian Analysis.** We computed the one-sided default Bayes factors for the replications[31], as the hypotheses tested in the replications are clearly one-sided. Hence we obtain the strength of evidence in favor of the hypothesis that stipulates an effect in the direction of the original experiment (where a default prior is assigned to the size of the effect, that is, a folded Cauchy distribution with scale 0.707) versus the null hypothesis that stipulates the effect to be absent. The default Bayes factor are shown in Fig. 3. Following Marsman et al.[32] the default Bayes factors in the figure are interpreted in terms of the evidence categories proposed by Jeffreys[33] (from extreme support for the null hypothesis to extreme support for the original hypothesis).

We furthermore computed the replication Bayes factor. The replication Bayes factor is similar to the default Bayes factor but uses the posterior distribution of effect size from the original experiment as the prior distribution of effect size under the alternative hypothesis for the evaluation of the replication study[34]. A replication Bayes factor above one favors the effect size observed in the original study and a replication Bayes factor below 1 favors the null hypothesis of no effect. The replication Bayes factor are shown in Supplementary Figure 3. The Replication Bayes factor exceeds 1, showing evidence in favor of the effect size observed in the original study for 12 (57.1%) studies (all of whom replicated according to our primary replication indicator). This evidence is strong to extreme for 9 (42.9%) studies and moderate for the remaining 3 studies. The default Bayes factor is below 1 for 9 (42.9%) studies showing evidence in support of the null hypothesis of a zero effect size over the effect size observed in the original study; this evidence is strong to extreme for 7 (33.3%) studies and moderate for 2 studies.

The replication Bayes factor yields similar results to the default Bayes factor with the exception of one study. For the Janssen et al.[3] study the Bayes factor shifts direction. The one-sided default Bayes factor for this study shows moderate support for an effect in the same direction as the original study, and the replication Bayes factor shows extreme support for a zero effect size over the effect size in the original study. This is consistent with an effect in the original direction for this replication, but with an effect size closer to the null effect than the effect size observed in the original study. The default Bayes factor tests the same thing as in our primary replication indicator and the test in the replication Bayes factor is similar to the test in the prediction interval approach (the prediction interval test if there is a significant difference between the replication effect size and the original effect size). The Janssen et al.[3] study has a significant effect in the original direction in our primary replication test consistent with the default Bayes factor. However, the replication result is within the 95% prediction interval, although close to the lower bound. The prediction interval indicator and the replication Bayes factor thus reach different conclusions. The Janssen et al.[3] study is the only study using a non-parametric Mann-Whitney test in our study. In a robustness test below we show that the prediction interval result for this study is sensitive towards if a $t$-test or a Mann-Whitney test is used as a basis of converting the original study result into a standardized effect size. With a

$t$-test the prediction interval result is in line with the replication Bayes factor result.

In addition to computing Bayes factors, we also estimated a Bayesian mixture model of the overall results[35,36]. The Bayesian mixture model provides an estimate of both the rate of true positives in the sample, and the relative effect size among true positive findings. The Bayesian mixture model assumes that each replication study originates from one of two components. The first component is the null hypothesis, according to which the expected effect size in a replication study is zero. The second component is the alternative hypothesis, according to which the expected effect size in a replication study equals a proportion of that from the original study. This is the effect size deflation factor for true positive studies – the extent to which replication studies yield effect sizes smaller than those obtained in the original studies for true positive studies. This estimate is important for determining the appropriate power to use in replication studies (i.e., the power has to take into account that effect sizes of true positive findings are likely to be inflated in original studies). All analyses were performed on the Fisher-transformed effect sizes $r$. An errors-in-variables mixture model was estimated to take into account the uncertainty about the reported effect sizes[37]. The results are shown in Supplementary Figure 5. A more detailed report about the estimation of the Bayes factors and the Bayesian mixture models are posted at OSF (https://osf.io/pfdyw/).

**Robustness Analysis.** In the *RPP*, the meta-analysis was only carried out for the subset of studies based on $t$-tests and one degree of freedom $F$-tests (which is identical to a $t$-test), as no standard errors of the correlation coefficients were estimated for studies using other tests[7]. We therefore carried out a robustness test based on $t$-tests of the treatment effect for all the 21 studies. For the five original studies using a $z$-test statistic, we re-analyzed the original data and the replication data with a $t$-test. We then based the estimated effect sizes on the $t$-tests in the meta-analytic estimates, the prediction interval estimates and the Small telescopes estimates.

Balafoutas and Sutter[4] and Gneezy et al.[5] used a $z$-test of the differences in proportions between two treatments. The data in these studies were re-analyzed using an independent samples $t$-test. This gives very similar results. For Balafoutas and Sutter[4] the $z$-value of the original study is 2.371 ($p = 0.018$) and the effect size is $r = 0.278$; with a $t$-test the

$t$-value is $t(70) = 2.436$ ($p = 0.017$) and the effect size is $r = 0.280$. In the replication the $z$-value is 2.285 ($p = 0.022$) and the effect size is $r = 0.146$; with a $t$-test the $t$-value is $t(241) = 2.300$ ($p = 0.022$) and the effect size is $r = 0.147$. The relative effect size of the replication based on the $z$-test is 0.527 and the relative effect size of the replication based on the $t$-test is 0.524. For Gneezy et al.[5] the $z$-value of the original study is 3.000 ($p = 0.003$) and the effect size is $r = 0.223$; with a $t$-test the $t$-value is $t(176) = 3.066$ ($p = 0.003$) and the effect size is $r = 0.225$. In the replication the $z$-value is 3.706 ($p < 0.001$) and the effect size is $r = 0182$; with a $t$-test the $t$-value is $t(405) = 3.761$ ($p < 0.001$) and the effect size is $r = 0.184$. The relative effect size of the replication based on the $z$-test is 0.818 and the relative effect size of the replication based on the $t$-test is 0.816.

Derex et al.[38] estimated a logistic regression with the probability of maintaining cultural diversity as a function of group size, and the group size coefficient was evaluated with a Wald test (equivalent to a $z$-test). The data in this study was re-analyzed with a linear probability model. The $z$-value of the group size variable coefficient in the original study is 4.037 ($p < 0.001$) and the effect size is $r = 0.525$; with a $t$-test in a linear probability model the $t$-value is $t(49) = 4.812$ ($p < 0.001$) and the effect size is $r = 0.566$. In the replication the $z$-value of the group size coefficient is 2.972 ($p = 0.003$) and the effect size is $r = 0.361$; with a $t$-test in a linear probability model the $t$-value is $t(63) = 3.625$ ($p < 0.001$) and the effect size is $r = 0.415$. The relative effect size of the replication based on the $z$-test is 0.687 and the relative effect size of the replication based on the $t$-test is 0.733.

Janssen et al.[3] used a Mann-Whitney test (yielding a $z$-value) to compare group earnings between two treatments and the data in this study was re-analyzed with an independent samples $t$-test. The $z$-value of the original study is 5.761 ($p < 0.001$) and the effect size is $r = 0.631$; with a $t$-test the $t$-value is $t(61) = 15.104$ ($p < 0.001$) and the effect size is $r = 0.888$. In the replication the $z$-value is 2.238 ($p = 0.025$) and the effect size is $r = 0.344$; with a $t$-test the $t$-value is $t(40) = 2.413$ ($p = 0.020$) and the effect size is $r = 0.356$. The relative effect size of the replication based on the $z$-test is 0.545 and the relative effect size of the replication based on the $t$-test is 0.401.

Rand et al.[9] estimated a Tobit model of the contribution in a public goods game as a function of a treatment dummy variable and the data in this study was re-analyzed with an independent samples $t$-test. The $z$-value of the treatment dummy variable coefficient in the original study is 2.617 ($p = 0.009$) and the effect size is $r = 0.1410$; with a $t$-test the $t$-value is $t(341) = 2.446$ ($p = 0.015$) and the effect size is $r = 0.131$. In Stage 1 of the replication data collection, the $z$-value of the treatment dummy variable coefficient is 0.904 ($p = 0.366$) and the effect size is $r = 0.028$; with a $t$-test the $t$-value is $t(1012) = 0.693$ ($p = 0.488$) and the effect size is $r = 0.022$. In Stage 2 of the replication, the $z$-value of the treatment dummy variable coefficient is 1.191 ($p = 0.234$) and the effect size is $r = 0.026$; with a $t$-test the $t$-value is $t(2134) = 1.006$ ($p = 0.315$) and the effect size is $r = 0.022$. The relative effect size of the replication based on the $z$-test is 0.183 and the relative effect size of the replication based on the $t$-test is 0.166.

The results of the robustness test are shown in Supplementary Figure 1 and Supplementary Figure 2. The results are overall very similar in the robustness test. However, in the prediction interval approach the conclusion about replication changes for two studies; but in different directions leaving the overall replication rate unchanged. For the Janssen et al.[3] study the replication effect size based on a $z$-test is within the prediction interval (although very close to the lower bound of the interval), but based on a $t$-test the replications falls below the prediction interval. This is driven by the $t$-test being even more strongly significant than the Mann-Whitney test for the original study. Overall there is a significant effect in the original direction for this study, but the effect is significantly smaller than in the original study. This is in line with Bayes factor results for this study; the default Bayes factor showed support for the original hypothesis over the null hypothesis, and the replication Bayes factor showed that the effect size is closer to the null effect size than the original effect size. For the study by Rand et al.[9], the replication effect size based on a $z$-test does not fall within the prediction interval (although very close to the lower bound of the interval), but based on a $t$-test the effect size is within the prediction interval. The difference in results for Rand et al. is small and it is close to the border of the prediction interval in both cases, but just outside the interval based on the $z$-test and just inside the interval based on the $t$-test. In the robustness test, 14 effects replicate (66.7%) according to the prediction interval approach. This is the same as for the main analysis.

The only other change in the robustness analysis is that the Rand et al.[9] study does not replicate in the meta-analysis. This is due to a small change in the meta-analytic $p$-value for this study from 0.038 to 0.066. In the robustness test of the meta-analysis 15 studies (71.4%) have a significant effect in the same direction as the original study. However, if the lower $p$-value threshold of 0.005 for statistically significant new findings suggested by Benjamin et al.[33] is applied, the conclusion for Rand et al.[9] in the meta-analysis is not affected in the robustness test and 13 studies or 61.9% still have a $p$-value $< 0.005$ in the meta-analysis.

The mean standardized effect size (correlation coefficient, $r$) of the replications in the robustness test is 0.252, compared to 0.473 in the original studies. The mean relative effect size of the replications is 45.7%. For the 13 studies that replicated, the mean relative effect size is 73.7%, and for the 8 studies that did not replicate, the mean relative effect size is 0.1%. These results are almost identical to the initial results.

## Implementation of prediction markets and surveys

We used both surveys and prediction markets to measure peer beliefs about replicability. Prediction markets can be utilized as a mechanism to aggregate private information and beliefs and have been successfully applied to make predictions in several fields[39–45]. Prediction markets and surveys were used to estimate peer beliefs about replication in a subset of the studies in the *RPP*[46], and for the replications in the *EERP*[8].

**Treatments.** We used two different treatments for eliciting peer beliefs. Participants were randomly assigned to these two treatments after signing up to participate in the study, and prior to filling out the survey. In both treatments we elicited beliefs with both a survey and with prediction markets for the 21 replication studies. In Treatment 1 we elicited beliefs about replicability in the first data collection (Stage 1). In Treatment 2 we elicited beliefs about replication in the first data collection (Stage 1) and in the first and second data collection pooled (Stage 2). In the prediction markets, shares could be traded whose value was determined by the actual outcome of the replication. Using two different treatments allowed us to implement a design that is as simple as possible (Treatment 1) and identical

to previous prediction market settings[8,46], as well as a more complex design to elicit predictions specific to the two-stage approach used in *SSRP* (Treatment 2). Comparison of the two treatments allowed us to investigate how robust forecasts from surveys and prediction markets are.

**Recruitment.** We sent invitations to participate in the survey and prediction markets to the Economic Science Association mailing list, the Society for Judgment and Decision Making e-mail list, the OSF e-mail list and the PsychMAP Facebook group. The invitation was also tweeted by Brian Nosek. The invitation contained a link to an online form where participants could sign up using their email address. A PhD degree or currently being a PhD student was a requirement for participating in the survey and prediction markets. The invitations to participate in the survey and prediction markets were e-mailed on October 17, 2016 and registrations closed on October 31. The survey was sent out to those who had registered by October 31 and the deadline for completing the survey was November 5. The prediction markets opened on November 7 and closed on November 21.

**Participants.** Initially, 397 individuals signed up to participate and were randomly assigned to one of the two treatments; 128 filled in the pre-market survey in Treatment 1 and 104 filled in the pre-market survey in Treatment 2; 114 participated on the prediction markets in Treatment 1 and 92 participated in the prediction market in Treatment 2. The number of traders active in each of the markets ranged from 36 to 80 in Treatment 1 and 18 to 68 in Treatment 2. Of the participants 6.8% did not work in academia (but had a PhD), 35.0% were PhD students, 35.0% were post-docs or assistant professors, 11.2% were lecturers or associate professors, and 11.7% were full professors. The average time spent in academia after obtaining the PhD (for the 80% who answered this question) was 6.0 years. 41.3% of the participants resided in Europe and 51.0% resided in North America. The most common core field of research was psychology (51.5%) followed by economics (40.3%).

**Information available to participants.** All participants had access to the replication reports for each replication (the version of the replication reports before the replications were conducted), and the references to the original papers. In the instructions to the survey and prediction markets, participants were also informed that the statistical power

was 90% to detect 75% of the original effect size in Stage 1, and 90% power to detect 50% of the original effect size in Stage 2 (and that the criteria for replication was a $p$-value $< 0.05$ in a two-sided test and an effect in the same direction as the original study). For each replication study, participants were informed about the hypothesis to be replicated, the $p$-value of the original result, and the sample size of both the original study and the replication.

**Elicitation of peer beliefs about replicability.** The pre-market survey (available at www.socialsciencesreplicationproject.com) was designed to elicit the same type of information as the prediction markets (i.e., the beliefs about replicability). Participants in the pre-market survey in Treatment 1 were asked to assess, for each replication study: (i) the likelihood that the hypothesis would be replicated in Stage 1 of the data collection; (ii) their stated expertise for the study the hypothesis was taken from. Participants in the pre-market survey in Treatment 2 were asked to assess, for each replication study: (i) the likelihood that the hypothesis would be replicated in Stage 1 of the data collection; (ii) the likelihood that the hypothesis would not be replicated in Stage 1, but would be replicated in Stage 2 with the pooled data; (iii) the likelihood that the hypothesis would not be replicated in Stage 1 or in Stage 2; (iv) their stated expertise for the study the hypothesis was taken from. Participants could also optionally answer a few demographic questions. The survey questions were not incentivized.

**Implementation of prediction markets.** To implement the prediction markets we used the same web-based trading platform as in the *EERP*[8], but adjusted the software for Treatment 2. There were two main views on the trading interface: (i) the market overview and (ii) the trading page. The market overview showed the 21 markets alongside summary information and a trade button for each market. The trading page was shown after clicking the trade button; at the trading page the participant could make investment decisions and view more detailed information about the market (see Supplementary Figure 6).

**Trading and market pricing.** In both treatments, the prediction markets participants were endowed with 100 Tokens. Once the markets opened, these Tokens could be used to trade shares in the markets. For each share held at market closing, participants received one Token if the corresponding outcome was realized and zero otherwise. Prices

for this type of share are typically interpreted as the predicted probability of the outcome to occur[47,48]; see Sonneman et al.[49] for lab evidence that averaged beliefs are close to prediction market prices.

In Treatment 1, participants could trade shares that paid one Token if a study replicated after Stage 1, as defined by a $p$-value $< 0.05$ in a two-sided test and an effect in the same direction as in the original study. Participants in this treatment could short-sell, which is equivalent to buying shares that pay one Token if the study was not replicated after Stage 1. Prices were determined by a market maker implementing a logarithmic market scoring rule[50] for a binary outcome, with a liquidity parameter of $b = 100$. As in previous studies[8,46], markets in this treatment opened at a price of 0.50 Tokens per share (for replication in Stage 1); trading is described in detail in Camerer et al.[8].

In Treatment 2, participants could trade three types of shares for each study, corresponding to replication in Stage 1 ($p$-value $< 0.05$ in a two-sided test and an effect in the same direction as in the original study based on the data collected in Stage 1), replication in Stage 2 ($p$-value $< 0.05$ in a double-sided test and an effect in the same direction as in the original study based on the combined data collected in Stage 1 and 2), and no replication. Short selling was not possible, but is not required as participants could directly trade shares on all three outcomes. Prices were determined by a market-maker implementing a logarithmic market scoring rule[50] for three mutually exclusive outcomes. Prices for infinitesimally small transactions of the three shares are given by $e^{(A_1+S_1)/b} \cdot Z^{-1}$, $e^{(A_2+S_2)/b} \cdot Z^{-1}$, and $e^{(A_3+S_3)/b} \cdot Z^{-1}$, with $Z = e^{(A_1+S_1)/b} + e^{(A_2+S_2)/b} + e^{(A_3+S_3)/b}$. $S_1$, $S_2$, and $S_3$ denote the market maker net sales of shares on replication at Stage 1, Stage 2, and no replication. As in Treatment 1, we used a liquidity parameter of $b = 100$; and the values for $A_1$, $A_2$, and $A_3$ were chosen such that markets opened at a price of 0.50 Tokens per share for replication in Stage 1, 0.25 Tokens per share for replication in Stage 2, and 0.25 Tokens per share for not replicating in either Stage 1 or Stage 2. These starting prices were chosen to be analogous to the ones used in Treatment 1.

Prices for finite transactions were obtained by integrating over the price function. The market maker ensures both that trades are always possible even when there is no other participant with whom to trade and that participants have incentives to invest according

to their beliefs[51]. In both settings, investment decisions for a market were made from the market's trading page. Participants could see the (approximate) price of a new share, the number of shares they currently held, and the number of Tokens their current position was worth if they liquidated their shares. Information about previous prices and aggregate positions was also displayed as graphs on the trading page. To make an adjustment to their current position, participants could choose either to increase or decrease their position by a number of Tokens of their choice.

See Supplementary Table 6 for data about trading volume on the prediction markets.

**Incentivisation.** The markets were resolved after all replication experiments were completed. If a replication was successful, shares held in the corresponding market were worth 1 Token. Tokens awarded as a result of holding shares were converted to USD at a 0.5 rate. Tokens that had not been invested in a market were not converted.

## Comparison of prediction market beliefs, survey beliefs, and replication outcomes

To compare the survey results to the prediction markets results we based the pre-market survey measure on the sample of individuals who participated on the prediction markets ($n = 114$ in Treatment 1 and $n = 92$ in Treatment 2). Prediction market beliefs and survey beliefs for Treatment 1 and Treatment 2 are shown in Supplementary Table 5. We analyzed the results separately for Treatment 1 and Treatment 2, and we focus on the results for Treatment 2 in the main text (as this measures beliefs about replication after Stage 2 using all the data on replication). All prediction market beliefs below refer to final trading prices on the study level.

### Treatment 1 results

Treatment 1 measures beliefs about replicability after Stage 1. Supplementary Figure 7 depicts the relationship between survey beliefs and prediction market beliefs from this treatment and how they relate to the replication outcome. Prediction market and survey beliefs

are strongly related and the Spearman correlation between the prediction market beliefs (final market prices) and the survey beliefs is 0.894 ($p < 0.001$, 95% CI = $[0.752, 0.956]$, $n = 21$). The range of predictions is 21.3% to 79.9% with a mean of 56.9% ($Mdn = 62.6\%$) in the prediction markets and 19.0% to 70.5% with a mean of 48.9% ($Mdn = 49.6\%$) in the survey. This can be compared to the observed replication rate of 57.1% after Stage 1; the prediction market and survey beliefs do not differ significantly from the observed replication rate in Stage 1 (Wilcoxon signed-ranks test, $z = 0.156$, $p = 0.876$, $n = 21$, for prediction market beliefs versus the observed replication rate and $z = 0.643$, $p = 0.520$, $n = 21$, for survey beliefs versus the observed replication rate). However, the prediction market beliefs are significantly higher than the survey beliefs (Wilcoxon signed-rank test, $z = 3.076$, $p = 0.002$, $n = 21$).

To evaluate if market beliefs and survey beliefs contain useful information for predicting replication outcomes, we estimated the Spearman correlation between beliefs and replication outcomes (i.e., the binary outcome whether a replication shows a statistically significant effect in the same direction as in the original study); we estimated these correlations both for the replication results after Stage 1 and the replication results after Stage 2 (basing the replication results on the maximum data). The Spearman correlation coefficient between beliefs and the replication outcome after Stage 1 is 0.509 ($p = 0.019$, 95% CI = $[0.098, 0.771]$, $n = 21$) for market beliefs and 0.540 ($p = 0.011$, 95% CI = $[0.142, 0.788]$, $n = 21$) for survey beliefs. The Spearman correlation coefficient between beliefs and the replication outcome after Stage 2 is 0.777 ($p < 0.001$, 95% CI = $[0.520, 0.905]$, $n = 21$) for market beliefs and 0.696 ($p < 0.001$, 95% CI = $[0.378, 0.867]$, $n = 21$) for survey beliefs. The mean absolute prediction error (the difference between beliefs and the replication outcome after Stage 1) is 39.7 percentage points for the prediction market beliefs and 42.3 percentage points for the survey beliefs, and they do not significantly differ from each other (Wilcoxon signed-rank test, $z = 1.442$, $p = 0.149$, $n = 21$).

Supplementary Table 7 summarizes the Spearman correlations between prediction market and survey beliefs and the other reproducibility indicators considered in the *SSRP*, and Supplementary Figure 8a plots the relationship between beliefs and the relative effect size of the replications after Stage 2 (i.e., basing the relative effect sizes on the maximum

data). The Spearman correlation coefficient between the prediction market beliefs and the relative effect size is 0.596 ($p = 0.004$, 95% CI = $[0.221, 0.817]$, $n = 21$), and the Spearman correlation coefficient between the survey beliefs and the relative effect size is 0.599 ($p = 0.004$, 95% CI = $[0.225, 0.819]$, $n = 21$).

### Treatment 2 results

Treatment 2 measures beliefs about replication after Stage 1, "Stage 2-Added" (the probability of not replicating in Stage 1, but replicating in Stage 2) and after Stage 2. In Fig. 4 we show the relationship between survey beliefs and prediction market beliefs after Stage 2, and how they relate to the replication outcome. The Spearman correlation between the prediction market beliefs and the survey beliefs about replication after Stage 1 is 0.860 ($p < 0.001$, 95% CI = $[0.681, 0.942]$, $n = 21$); the range of predictions is 14.8% to 83.4% with a mean of 45.5% ($Mdn = 46.0\%$) in the prediction markets and 12.9% to 54.4% with a mean of 36.6% ($Mdn = 34.1\%$) in the survey. The Spearman correlation between the prediction market beliefs and the survey beliefs about "Stage 2-Added-Replication" is 0.197 ($p = 0.391$, 95% CI = $[-0.256, 0.580]$, $n = 21$); the range of predictions is 6.0% to 31.3% with a mean of 17.9% ($Mdn = 19.4\%$) in the prediction markets and 14.8% to 29.7% with a mean of 24.0% ($Mdn = 24.7\%$) in the survey. The Spearman correlation between the prediction market beliefs and the survey beliefs about replication after Stage 2 is 0.845 ($p < 0.001$, 95% CI = $[0.652, 0.936]$, $n = 21$). The range of predictions is 23.1% to 95.5% with a mean of 63.4% ($Mdn = 62.7\%$) in the prediction markets and 27.8% to 81.5% with a mean of 60.6% ($Mdn = 60.3\%$) in the survey.

This can be compared to the observed replication rate of 57.1% after Stage 1, 4.8% in "Stage 2-Added" (including the Ackerman et al.[28] result in Stage 2, even though this study should not have been included in Stage 2 according to the prediction market protocol; not including the Ackerman et al.[28] result in Stage 2 would increase the number by 4.8 percentage points), and 61.9% after Stage 2 (including the Ackerman et al.[28] result in Stage 2, even though this study should not have been included in Stage 2 according to the prediction market protocol; not including the Ackerman et al.[28] result in Stage 2 would increase the replication rate by 4.8 percentage points). The test results for comparing

survey beliefs and prediction market beliefs to the observed replication rate are: prediction market beliefs about replication after Stage 1 versus observed (Wilcoxon signed-ranks test, $z = 1.130$, $p = 0.259$, $n = 21$) and survey beliefs about replication after Stage 1 versus observed (Wilcoxon signed-rank test, $z = 2.450$, $p = 0.014$, $n = 21$); "Stage 2-Added" prediction market beliefs versus observed (Wilcoxon signed-ranks test, $z = 0.866$, $p = 0.386$, $n = 10$) and "Stage 2-Added" survey beliefs versus observed (Wilcoxon signed-rank test, $z = 0.866$, $p = 0.386$, $n = 10$); prediction market beliefs about replication after Stage 2 versus observed (Wilcoxon signed-ranks test, $z = 0.469$, $p = 0.876$, $n = 21$) and survey beliefs about replication after Stage 2 versus observed (Wilcoxon signed-rank test, $z = 0.261$, $p = 0.794$, $n = 21$).

The Spearman correlation coefficient between beliefs about replicating after Stage 1 and the replication outcome after Stage 1 is 0.429 ($p = 0.052$, 95% CI $= [-0.003, 0.726]$, $n = 21$) for prediction market beliefs and 0.556 ($p = 0.009$, 95% CI $= [0.164, 0.797]$, $n = 21$) for survey beliefs. The mean absolute prediction error is 41.4 percentage points for the prediction market beliefs and 44.4 percentage points for the survey beliefs (Wilcoxon signed-rank test, $z = 0.643$, $p = 0.520$, $n = 21$). The Spearman correlation coefficient between beliefs of replicating after Stage 2 and the replication outcome after Stage 2 is 0.842 ($p < 0.001$, 95% CI $= [0.645, 0.934]$, $n = 21$) for prediction market beliefs and 0.761 ($p = 0.001$, 95% CI $= [0.491, 0.898]$, $n = 21$) for survey beliefs. The mean absolute prediction error is 0.303 for the prediction market beliefs and 0.348 for the survey beliefs (Wilcoxon signed-rank test, $z = 2.068$, $p = 0.039$, $n = 21$).

We also compare the estimated beliefs of replication after Stage 1 between Treatment 1 and Treatment 2. The estimated beliefs about replication was significantly higher in Treatment 1 for both prediction markets beliefs (Wilcoxon signed-rank test, $z = 3.667$, $p < 0.001$, $n = 21$) and survey beliefs (Wilcoxon signed-rank test, $z = 4.015$, $p < 0.001$, $n = 21$). This could be explained by the existence of a framing effect in Treatment 2; by asking subjects about beliefs of three events for each replication they lower their estimates of replication in Stage 1 and increases the added replication in Stage 2. However, the Spearman correlation between Treatments 1 and 2 for beliefs of replicating after Stage 1 is high (0.895, $p < 0.001$, 95% CI $= [0.755, 0.957]$, $n = 21$ for prediction market beliefs and 0.983, $p < 0.001$, 95% CI $= [0.958, 0.993]$, $n = 21$ for survey beliefs) showing a similar

ordering of the 21 replications in the two treatments consistent with a high test-retest reliability of both prediction market and survey beliefs. We also estimate the Spearman correlation between beliefs about replicating after Stage 2 in Treatment 2, and beliefs about replicating after Stage 1 in Treatment 1. These correlations are also high (0.890 ($p < 0.001$, 95% CI = $[0.743, 0.955]$, $n = 21$) for prediction market beliefs and 0.981 ($p < 0.001$, 95% CI = $[0.952, 0.992]$, $n = 21$) for survey beliefs) again supporting a high test-retest reliability of measuring prediction market and survey beliefs.

Supplementary Table 7 summarizes the Spearman correlations between prediction market and survey beliefs and the other reproducibility indicators considered in the *SSRP*; Supplementary Figure 8b plots the relationship between beliefs and the relative effect size of the replications. The Spearman correlation coefficient between prediction market beliefs about replicating after Stage 2 and the relative effect size is 0.642 ($p = 0.002$, 95% CI = $[0.290, 0.840]$, $n = 21$), and the Spearman correlation coefficient between survey beliefs about replicating after Stage 2 and the relative effect size is 0.621 ($p = 0.003$, 95% CI = $[0.258, 0.830]$, $n = 21$).

## Comparison of reproducibility indicators to experimental economics and psychological sciences

Supplementary Figure 9 compares the results for our two main replicability indicators (significant effect in the same direction as the original study and the relative effect size) to the results for psychological sciences in the *RPP*[7] and experimental economics in the *EERP*[8]. This comparison is based on the replication results after Stage 2. The results for the *RPP* study are the same ones as presented in the *EERP* paper[8] and they were taken from the published replication results[7]. The *RPP* did not directly report the relative effect size of the replication, but instead used the "effect size difference" as a reproducibility indicator. The "effect size difference" was estimated as the absolute difference in the standardized effect size ($r$) between the original study and the replication study. We prefer to use the relative effect size (the ratio between the standardized effect size ($r$) of the replication and the standardized effect size ($r$) of the original study). The reason for

this is the lack of comparability of the standardized effect sizes between our 21 studies as discussed in Section 2 above. We used the same relative effect size measure for the *RPP* as estimated by Camerer et al.[8]; they downloaded the posted effect size data from the *RPP* and estimated the relative replication effect size for each study.

## Results and data for the individual studies and markets

The hypotheses as described to the participants on the prediction markets in each of the 21 studies are shown in Supplementary Table 1. Supplementary Table 3 (Stage 1) and Supplementary Table 4 (Stage 2) summarize detailed replication results for the 21 studies. In Supplementary Table 5 we present the prediction markets beliefs and the survey beliefs for each of the 21 studies. Additional prediction market data for both treatments are shown in Supplementary Table 6. Supplementary Table 7 depicts a correlation matrix (Spearman rank correlations) for the reproducibility indicators (including the prediction market and survey beliefs in Treatment 1 and Treatment 2) and two original study characteristics ($p$-value and original sample size). The correlation between the original $p$-value and the the statistical significance criterion of replication is discussed in the main text. The *EERP*[8] also found a sizable positive correlation between the original sample size and replicability, but this was not found in the *RPP*[7]. We therefore included the original sample size in the correlation matrix as well. As some analyses in the *SSRP* are performed on the group rather than the individual level, we correlate replication success with both the number of observations and the number of participants as two indicators for sample size. The correlation with the statistical significance criterion are negative but not significant ($-0.292$, $p = 0.199$, 95% CI $= [-0.643, 0.160]$, $n = 21$ for the number of observations and $-0.057$, $p = 0.807$, 95% CI $= [-0.477, 0.384]$, $n = 21$ for the number of participants) consistent with the findings from the *RPP*.

The number of observations and participants in each replication study are reported in Supplementary Table 3 and Supplementary Table 4. In some of the replications we also collected data about gender and age, but this was not collected in all replications (the replication data collections were based on the data collected in the original study). Below

we list the number of men and women and the mean age of the participants for the replication studies where this information is available (the data is for the pooled Stage 1 and Stage 2 data for the replications that proceed to Stage 2).

- Ackerman et al.[28] replication: 187 men and 408 women (4 participants refused to provide gender information); mean age = 29.2 years.
- Aviezer et al.[17] replication: No data on gender or age collected.
- Balafoutas and Sutter[4] replication: 243 men and 243 women; no data on age collected.
- Derex et al.[38] replication: 482 men, mean age = 23.7 years.
- Duncan et al.[12] replication: No data on gender or age collected.
- Gervais and Norenzayan[13] replication:196 men and 332 women (3 participants refused to provide gender information); mean age = 19.0 years.
- Gneezy et al.[5] replication: 160 men and 267 women, mean age = 20.3 years.
- Hauser et al.[26] replication: 47 men and 63 women, mean age = 36.5 years.
- Janssen et al.[3] replication: No data on gender or age collected.
- Karpicke and Blunt[19] replication: 23 men and 26 women, mean age = 19.1 years.
- Kidd and Castano[24] replication: 328 men and 386 women, mean age = 35.2 years.
- Kovacs et al.[20] replication: 46 men and 49 women; no data on age collected.
- Lee and Schwarz[25] replication: 97 men and 166 women (23 refused to provide information on gender); mean age = 19.0 years.
- Morewedge et al.[6] replication: 36 men and 53 women, mean age = 23.0 years.
- Nishi et al.[16] replication: No data on gender or age collected.
- Pyc and Rawson[52] replication: No data on gender or age collected.
- Ramirez and Beilock[22] replication: 42 men and 89 women, mean age = 19.2 years.
- Rand et al.[9] replication: 1080 men and 1056 women, mean age = 36.8 years.
- Shah et al.[27] replication: 280 men and 339 women, mean age = 36.7 years.

- Sparrow et al.[23] replication: No data on gender or age collected.
- Wilson et al.[14] replication: 10 men and 29 women, mean age = 20.3 years.

Below we also list any "unplanned protocol deviations" for each of the replications (based on this section of the "Replication Reports"). For the replications not included below, no "unplanned protocol deviations" occurred (we do not include the somewhat larger than planned sample sizes in some replications among the "unplanned protocol deviations" below).

**Ackerman et al.[28]:** An error occurred during analysis of the First Data Collection analysis (90% power to detect 75% of the original effect size), in which the preregistered analysis script was run with most, but not all, of the participants in the sample. The observed $p$-value in that analysis did not meet criteria for concluding data collection ($p < 0.05$). As such, we initiated the $2^{nd}$ round of data collection being run (90% power to detect 50% of the original effect size). The analysis error was discovered when the $2^{nd}$ round of data was nearly complete. We decided to finish the 2nd data collection and report all results for completeness. Additionally, the only exclusion criteria specified in the analysis section of our pre-data collection replication report encompassed, "sitting down or resting the clipboard on a surface." However, to maintain the integrity of the sample following data collection, we have employed additional criteria to exclude participants with a reported age below 18 years, as well as participants that reported previous knowledge of the experiment or correctly guessed the purpose of the study before or during the debriefing period. When all participants excluded due to values below 18 for age are included, those assigned to the heavier clipboard ($n = 299$) still rated the resume as similar overall compared to those assigned to the lighter clipboard ($n = 307$): Heavy clipboard $M = 5.84$, $SD = 0.85$; Light clipboard $M = 5.73$, $SD = 0.89$; $t(604) = 1.4548$, $p = 0.146$, $d = 0.1182$ $[-0.0415, 0.2779]$, $r = 0.0591$ $[-0.0208, 0.1382]$.

When including all the participants who reported previous knowledge of the experiment, those assigned to the heavier clipboard ($n = 300$) still rated the resume as similar moverall compared to those assigned to the lighter clipboard ($n = 305$): Heavy clipboard $M = 5.84$, $SD = 0.85$; Light clipboard $M = 5.73$, $SD = 0.89$; $t(603) = 1.4893$, $p = 0.137$, $d = 0.1211$ $[-0.0387, 0.2809]$, $r = 0.0605$ $[-0.0194, 0.1397]$.

Finally, when including all the participants except those who had missing values and those who were excluded based on the original exclusion criteria, those assigned to the heavier clipboard ($n = 303$) still rated the resume as similar overall compared to those assigned to the lighter clipboard ($n = 309$): Heavy clipboard $M = 5.84$, $SD = 0.85$; Light clipboard $M = 5.74$, $SD = 0.89$; $t(610) = 1.4093$, $p = 0.159$, $d = 0.1139$ $[-0.0450, 0.2728]$, $r = 0.0570$ $[-0.0225, 0.1366]$.

**Derex et al.[38]:** Due to difficulties in recruiting, the show-up fee was raised from s$5.00 to s$10.00, and the average performance-based payment was raised from s$10.00 to s$20.00.

**Hauser et al.[26]:** In the recruitment of subjects from AMT we only recruited subjects from the US, used HIT approval rate greater or equal to 85%, and number of HITs approved greater than or equal to 100. These criteria were suggested by the original authors and decided prior to starting the data collection.

**Janssen et al.[3]:** In the replication experiment, it turns out that for the *NCP-C* treatment the condition switch (from *NCP* to *C*) happened after the fourth period. Due to this incorrect condition switching, the *NCP-C* treatment in the replication experiment includes 4 rounds of the *NCP* condition followed by 2 rounds of the *C* condition. This does not affect the findings of our replication because as planned our analysis only uses data of the first 3 rounds of the treatments and the condition switch is by design unexpected by the participants.

**Kidd and Castano[24]:** We only used US AMT workers with a HIT approval rate of at least 95% in line with the original study, but this was not specified in the pre-replication version of this Replication Report. These criteria were decided upon together with the original authors before starting the data collection.

Two unplanned exclusion criteria were added. First, we excluded everyone who had at least one missing answer out of the 36 questions on the RMET test (31 in first data collection, 53 in second). This is because the dependent variable RMET score can not be constructed for a subject with one or more missing values on the RMET score. Since the original authors had no missing values in this outcome variable, they did not have to consider this exclusion. This exclusion criteria was decided together with the original

authors before starting the data collection.

Second, we excluded all participants with a negative score on the author recognition test (2 in first data collection, 6 in the second). A negative score here indicates that the subject had more incorrect guesses of authors than correct guesses. Since the analysis uses a square root transformation of this variable, these subjects could not be used in the analysis. This exclusion criteria was implemented after starting the data collection, as we had not foreseen the possibility of a negative score on the author recognition test.

Furthermore, we changed the definition of one of our exclusion criteria for our main replication result. We planned to exclude all participants with less than 30 seconds reading time as this was suggested by the original authors. However, to apply this threshold of 30 seconds reading time the original authors also wanted us to standardize the page length of the reading texts (which had not been done in the original study). However, we did not standardize the page length as we in the communication with the original authors did not understand that we should standardize the page length to apply the 30 seconds reading time threshold. After the data had been collected we therefore together with the original authors decided not to exclude participants with less than 30 seconds reading time, but to exclude participants with 0 reading time. Note that this implies that this exclusion criteria is now the same as in the original study (as the original study excluded participants with 0 reading time, and did not standardize the page length).

Note also that some observations where participants initiate the study are automatically excluded as some participants did not proceed further than giving consent and some participants did not proceed further than the instructions.

**Kovacs et al.[20]:** In the original study the tests were conducted in a 3m×3m sound-attenuated booth using Psyscope X on an Apple PowerBook. Due to difficulties in finding exactly the same experimental venue and materials, in the replication experiment the tests were carried out in a 3m×4m breakout room with participants wearing a sound-attenuated device (earmuffs with a noise reduction rating of 31 decibels) using Psyscope X on a 13-inch Apple MacBook Air.

**Lee and Schwarz[25]:** During our correspondence with the original authors, it was indicated that the original experimenter ensured that the presentation order of albums was

different in the pre- vs. post-manipulation evaluation, which was necessary to minimize the likelihood that participants provided their post-manipulation rankings by simply retrieving their memory of pre-manipulation rankings. Specifically, in the original experiment, participants provided their pre-manipulation rankings by writing down the album titles in the order they wanted, and then during filler tasks, the experimenter (in a different room) would prepare a different form by listing the albums in alphabetical order of the artist names (rather than album titles). In effect, the presentation order of albums would differ in the pre- vs. post-manipulation evaluation.

We noted this step in our "Procedure Script" prior to the initiation of data collection. However, we inadvertently neglected the alphabetization step when filling in the "Secondary Album Ranking" form for participants. Instead, the presentation order was the same as what was in the pre-manipulation evaluation.

We conducted some exploratory analysis on the possible influence of episodic memory on the presence of the effect of post-decisional dissonance. We separated participants who had transcribed their selected albums in ascending order by rank from those who did not. The former would be more likely to remember their album rankings by using the rank-ordering on the page as a cue. As such, if we remove these participants, we might observe some evidence for a difference between the examining and hand-washing conditions for participants that did not have this memory cue. Out of the 285 participant responses (not including the five exclusions from the pooled study sample) for the pre-manipulation, 57.5% ($n = 164$) of participants transcribed their 10 selected albums in the order in which they were ranked, ascending 1–10 directly down the form. The remaining 42.5% ($n = 121$) of participants did not.

Among the 121 participants that did not list the albums in rank order, we observed a large dissonance effect ($d = 1.13$) of approximately the same size as for the full sample ($d = 1.05$). Participants had a lower rank difference before the choice ($M = 0.39$, $SD = 0.93$) than after the choice ($M = 1.95$, $SD = 1.73$); $t(120) = 8.8162$, $p < 0.001$, $d = 1.13$ $[0.85, 1.40]$. The focal hypothesis, however, was not supported by the data. The rank difference between the chosen and rejected CDs did not differ between the participants in the examining conditions ($M = 1.60$, $SD = 1.65$) and the hand-washing condition

($M = 1.52$, $SD = 2.19$), $F(1, 119) = 0.0500$, $p = 0.8234$. The direction of the effect, though insignificant, was the same as the equivalent test in the original study.

**Morewedge et al.[6]:** As in the original study, observations more extreme than 2.5 standard deviations from the overall mean were dropped from the analysis. The overall mean of consumed M&M's was $M = 10.29$ grams ($SD = 8.99$). 6 participants consumed more than $M + 2.5SD = 32.76$ grams of M&M's and one participant did not consume any M&M's at all such that 7 observations were excluded from the analysis. Since no participant refused to consume any M&M's in the original study, the original authors did not have to consider this exclusion. However, this exclusion criteria was decided on together with the original authors before starting the data collection.

**Nishi et al.[16]:** In line with the original study we did not restrict the sample to only American Turkers, but this was not specified in the pre-replication version of the Replication Report. We only used AMT workers with a HIT approval rate of at least 95%, but this was not specified in the pre-replication version of this Replication Report (and it is unclear if the original study used any restrictions on the HIT approval rate for participating in the study).

One session was by mistake conducted with only 10 subjects. When the number of subjects finishing the training rounds did not reach at least 13, the attempted session was supposed to be canceled. By mistake, this was not done for one session in the visible treatment, but rather the game was played with only 10 subjects. The mean Gini over the 10 rounds was 0.1764 in this session which is quite close to the mean of 0.1690 for all the sessions in the visible treatment group. In spite of the inclusion of this session with only 10 subjects, the average group size (16.5) was the same for both the visible and the invisible treatment.

We planned to carry out the statistical test of the replication result using the model with multiway clustering on session and round as this was used to test the hypothesis in the original study. However, as this model produced implausible standard errors in the replication we instead used a model with only clustering on the session level as the main replication result. This does not affect the conclusion about whether the original study replicates.

It was challenging to carry out this replication. The original study was programmed in a version of the program Breadboard, which is no longer supported and the original authors did not provide the source code for the original version of the experiment. We thus had to program the experiment from the beginning using the new version of Breadboard. We experienced several problems with using Breadboard, and the program did not always function correctly. During the data collection the experimental software broke down 21 times and in these cases the session had to be restarted with new subjects.

We relied on the screenshots provided in the Supplementary Information to program the experiment as closely as possible to the original study. But as we did not have access to the source code of the original experiment and the original authors provided much guidance initially but were not able to reply to all subsequent detailed queries, we had to decide on a number of issues that were not clear from the original paper or Supplementary Information (these issues are listed at the end of the replication report).

**Pyc and Rawson[52]:** Contrary to our initial expectations, data for this replication was collected largely during the summer. While the original study used undergraduate participants exclusively, this replication used undergraduate and graduate students, as recruitment rates were low during the summer.

A second deviation from the protocol was the additional robustness check conducted by analyzing the manually coded responses. We did not anticipate the issue of potential misclassification of successful recalls due to typos before data-collection (it was not mentioned in the original publication).

**Ramirez and Beilock[22]:** We inadvertently oversampled high-pressure conditions and undersampled low-pressure conditions in the second data collection – we targeted 33/cell for each of the four conditions, and instead ended up with 25 in the low-pressure control, 27 in the low-pressure expressive writing, 45 in the high-pressure control, and 34 in the high-pressure expressive writing conditions at the conclusion of the Spring 2017 academic semester.

All conditions for both the first and second rounds of collection were randomly assigned using a random number generator (found at Random.org), to generate numbers 1–4, with each digit assigned to one of the 4 experimental conditions. This created a random im-

balance between the two high pressure conditions over the course of the pooled first and second round collections. Additionally, prioritization was placed on reaching the target sample for the main high pressure conditions used in the focal analysis during the Spring 2017 collection period, which led to an imbalance in the manipulation-check low pressure conditions. The sample that we had already collected (79 in the high-pressure conditions, and 52 in the low-pressure conditions) gave us better than 99.9% power to detect the original manipulation-check effect size of $d = 0.99$, and 79% power to detect 50% of that effect size. Also, because we oversampled the high-pressure conditions, we had more power to detect the focal tests than prespecified.

**Rand et al.[9]:** We only used AMT workers with a HIT approval rate of at least 95% in line with the original study, but this was not specified in the pre-replication version of this Replication Report. This criteria was decided upon together with the original authors before starting the data collection.

**Shah et al.[27]:** We ran a test round with 10 observations to test that the program worked as intended before getting some more feedback from the authors. These 10 observations were not included in the analysis as we received new information about the MTurk recruitment criteria used in the original study after collecting these observations. We updated the HIT to only accept users from the US as was done in the original study, and we added a requirement for a HIT success rate of at least 95% (this was decided in agreement with the original authors).

One observation in the first data collection and 9 in the second were incomplete and could not be used, because the program had not recorded their treatment group or subject id. It is possible that these users never finished the task, or maybe the software did not work properly on their platform. While we cannot link these observations to specific users, a few users did report minor technical problems.

**Wilson et al.[14]:** Some of the participants (10/39) self-reported being non-psychology majors, even though the recruitment details emphasized that only psychology majors were eligible to participate in the study. Apart from that, the replication study was conducted exactly the way outlined above, without additional deviations from protocol.

## Supplementary References


1. Falk, A. & Szech, N. Morals and markets. *Science* **340,** 707–711 (2013). *DOI: 10.1126/science.1231566*

2. Gelstein, S. *et al.* Human tears contain a chemosignal. *Science* **331,** 226–230 (2011). *DOI: 10.1126/science.1198331*

3. Janssen, M. A., Holahan, R., Lee, A. & Ostrom, E. Lab experiments for the study of social-ecological systems. *Science* **328,** 613–617 (2010). *DOI: 10.1126/science.1183532*

4. Balafoutas, L. & Sutter, M. Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science* **335,** 579–582 (2012). *DOI: 10.1126/science.1211180*

5. Gneezy, U., Keenan, E. A. & Gneezy, A. Avoiding overhead aversion in charity. *Science* **346,** 632–635 (2014). *DOI: 10.1126/science.1253932*

6. Morewedge, C. K., Huh, Y. E. & Vosgerau, J. Thought for food: imagined consumption reduces actual consumption. *Science* **330,** 1530–1533 (2010). *DOI: 10.1126/science.1195701*

7. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349,** aac4716 (2015). *DOI: 10.1126/science.aac4716*

8. Camerer, C. F. *et al.* Evaluating replicability of laboratory experiments in economics. *Science* **351,** 1433–1436 (2016). *DOI: 10.1126/science.aaf0918*

9. Rand, D. G., Greene, J. D. & Nowak, M. A. Spontaneous giving and calculated greed. *Nature* **489,** 427–430 (2012). *DOI: 10.1038/nature11467*

10. Bouwmeester, S. et al. Registered Replication Report Rand, Greene & Nowak (2012). *Perspect. Psychol. Sci.* **12,** 527–542 (2017). *DOI: 10.1177/1745691617693624*

11. Tinghög, G. et al. Intuition and cooperation reconsidered. *Nature* **498**, E1–E2 (2013). *DOI: 10.1038/nature12194*

12. Duncan, K., Sadanand, A. & Davachi, L. Memory's penumbra: episodic memory decisions induce lingering mnemonic biases. *Science* **337,** 485–487 (2012). *DOI: 10.1126/science.1221936*

13. Gervais, W. M. & Norenzayan, A. Analytic thinking promotes religious disbelief. *Science* **336,** 493–496 (2012). *DOI: 10.1126/science.1215647*

14. Wilson, T. D. *et al.* Just think: the challenges of the disengaged mind. *Science* **345,** 75–77 (2014). *DOI: 10.1126/science.1250830*

15. Mani, A., Mullainathan, S., Shafir, E. & Zhao, J. Poverty impedes cognitive function. *Science* **341,** 976–980 (2013). *DOI: 10.1126/science.1238041*

16. Nishi, A., Shirado, H., Rand, D. G. & Christakis, N. A. Inequality and visibility of wealth in experimental social networks. *Nature* **526,** 426–429 (2015). *DOI: 10.1038/nature15392*

17. Aviezer, H., Trope, Y & Todorov, A. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* **338,** 1225–1229 (2012). *DOI: 10.1126/science.1224313*

18. Chen, D. L., Schonger, M. & Wickens, C. oTree—An open-source platform for laboratory, online, and field experiments. *J. Behav. Exp. Finance* **9,** 88–97 (2016). *DOI: 10.1016/j.jbef.2015.12.001*

19. Karpicke, J. D. & Blunt, J. R. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science* **331,** 772–775 (2011). *DOI: 10.1126/science.1199327*

20. 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). *DOI: 10.1126/science.1190792*

21. Fischbacher, U. z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* **10,** 171–178 (2007). *DOI: 10.1007/s10683-006-9159-4*

22. Ramirez, G. & Beilock, S. L. Writing about testing worries boosts exam performance in the classroom. *Science* **331,** 211–213 (2011). *DOI: 10.1126/science.1199427*

23. Sparrow, B., Liu, J. & Wegner, D. M. Google effects on memory: cognitive consequences of having information at our fingertips. *Science* **333,** 776–778 (2011). *DOI: 10.1126/science.1207745*

24. Kidd, D. C. & Castano, E. Reading literary fiction improves theory of mind. *Science* **342,** 377–380 (2013). *DOI: 10.1126/science.1239918*

25. Lee, S. W. S. & Schwarz, N. Washing away postdecisional dissonance. *Science* **328,** 709 (2010). *DOI: 10.1126/science.1186799*

26. Hauser, O. P., Rand, D. G., Peysakhovich, A. & Nowak, M. A. Cooperating with the future. *Nature* **511,** 220–223 (2014). *DOI: 10.1038/nature13530*

27. Shah, A. K., Mullainathan, S. & Shafir, E. Some consequences of having too little. *Science* **338,** 682–685 (2012). *DOI: 10.1126/science.1222426*

28. Ackerman, J. M., Nocera, C. C. & Bargh, J. A. Incidental haptic sensations influence social judgments and decisions. *Science* **328,** 1712–1715 (2010). *DOI: 10.1126/science.1189993*

29. 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). *DOI: 10.1177/1745691616646366*

30. Simonsohn, U. Small Telescopes: detectability and the evaluation of replication results. *Psychol. Sci.* **26,** 559–569 (2015). *DOI: 10.1177/0956797614567341*

31. Wagenmakers, E.-J. et al. Bayesian inference for psychology. Part II: Example applications with JASP. *Psychon. Bull. Rev.*, in press (2017). *DOI: 10.3758/s13423-017-1323-7*

32. Marsman, M. *et al.* A Bayesian bird's eye view of 'replications of important results in social psychology'. *R. Soc. Open Sci.* **4,** 160426 (2017). *DOI: 10.1098/rsos.160426*

33. Jeffreys, H. *Theory of probability*, (Oxford University Press, Oxford, UK, ed. 3, 1961). *ISBN: 9780198503682*

34. Verhagen, J. & Wagenmakers, E.-J. Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* **143,** 1457–75 (2014). *DOI: 10.1037/a0036731*

35. Lee, M. D. & Wagenmakers, E.-J. *Bayesian cognitive modeling: a practical course.* (Cambridge University press, Cambridge, UK, 2013). *ISBN: 9781107603578*

36. Frühwirth-Schnatter, S. *Finite mixture and Markov switching models.* (Springer, New York, 2006). *ISBN: 978-0-387-35768-3*

37. Matzke, D. *et al.* Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra: Psychol.* **3,** 25 (2017). *DOI: 10.1525/collabra.78*

38. 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). *DOI: 10.1038/nature12774*

39. Benjamin, D., *et al.* Redefine statistical significance. *Nat. Hum. Behav.* **1** (2017). *DOI: 10.1038/s41562-017-0189-z*

40. Arrow, K. J. *et al.* The promise of prediction markets. *Science* **320,** 877 (2008). *DOI: 10.1126/science.1157679*

41. Hanson, R. Could Gambling Save Science? Encouraging an Honest Consensus. *Soc. Epistemology* **9,** 3–33 (1995). *DOI: 10.1080/02691729508578768*

42. Wolfers, J. & Zitzewitz, E. Prediction Markets. *J. Econ. Perspect.* **18,** 107–26 (2004). *DOI: 10.1257/0895330041371321*

43. Tziralis, G. & Tatsiopoulos, I. Prediction Markets: An Extended Literature Review. *J. Pred. Mark.* **1,** 75–91 (2007). *DOI: 10.5750/jpm.v1i1.421*

44. Berg, J., Forsythe, R., Nelson, F. & Rietz, T. Results from a Dozen Years of Election Futures Markets Research. *Handbook of Experimental Economics Results* **1,** 742–51 (2008). *DOI: 10.1016/S1574-0722(07)00080-7*

45. Horn, C. F., Ivens, B. S., Ohneberg, M. & Brem, A. Prediction Markets – A Literature Review. *J. Pred. Mark.* **8,** 89–126 (2014). *DOI: 10.5750/jpm.v8i2.889*

46. Dreber, A, *et al.* Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl Acad. Sci. U.S.A.* **112,** 15343–15347 (2015). *DOI: 10.1073/pnas.1516179112*

47. Manski, C. F. Interpreting the Predictions of Prediction Markets. *Econ. Lett.* **91,** 425–29 (2006). *DOI: 10.1016/j.econlet.2006.01.004*

48. Wolfers, J. & Zitzewitz, E. Interpreting Prediction Market Prices as Probabilities (Working Paper No. 12200, National Bureau of Economic Research, 2006). *DOI: 10.3386/w12200*

49. Sonneman, U, Camerer, C. F., Fox, C. R. & Langer, T. How psychological framing affects economic market prices in the lab and field. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 11779–84 (2013). *DOI: 10.1073/pnas.1206326110*

50. Hanson, R. Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation. *J. Pred. Mark.* **1,** 3–15 (2007).

51. Chen, Y. *Markets as an information aggregation mechanism for decision support*, thesis, School of Information Sciences and Technology, Pennsylvania State University, State College, PA (2005).

52. Pyc, M. A. & Rawson, K. A. Why testing improves memory: mediator effectiveness hypothesis. *Science* **330,** 335 (2010). *DOI: 10.1126/science.1191465*

# Supplementary Tables

**Supplementary Table 1. Hypotheses for the 21 replication studies.**

| Study | Hypothesis |
|---|---|
| Ackerman et al. (2010), Science | Participants that evaluate a resume while using a heavier clipboard will rate the resume as better overall compared to the participants that evaluate the resume while using a lighter clipboard. The original study used $F$-test for a two condition comparison, $p < 0.05$. Original test statistics: Heavy Condition: $N = 26$, $M = 5.80$, $SD = 0.76$; Light Condition: $N = 28$, $M = 5.38$, $SD = 0.79$. $F(1, 52) = 4.08$, $p = 0.049$. If there were no covariates in the model, we will convert the $F$ to $t$ for comparison with the replication tests. |
| Aviezer et al. (2012), Science | The body context is diagnostic for the affective valence of the situation during peak intensity moments (tests the hypothesis of a higher mean valence rating of winning bodies versus losing bodies in the 'body treatment' in Experiment 1; within subjects variation, paired $t$-test, $t(14) = 13.07$, $p < 0.0001$, p. 1226 and Fig. 1c). |
| Balafoutas and Sutter (2012), Science | With preferential treatment of women – i.e., each woman's performance is automatically increased by one unit in the competition – more women will choose to compete (a comparison of the fraction of women who chose the tournament scheme rather than the piece rate scheme in the 'preferential treatment one (PT1)' versus the 'control treatment (CTR)'; $\chi^2(1) = 5.62$, $p = 0.018$, p. 580). (This hypothesis was picked by lottery instead of comparing PT2 to CTR; $\chi^2(1) = 10.89$, $p = 0.001$, p. 580). |
| Derex et al. (2013), Nature | The probability of maintaining cultural diversity (that is, observing both tasks in the group) increases with group size; $\chi^2(1) = 16.3$, the $p$-value $< 0.0001$ (exact 0.000054) (p. 389; measured at the group level with group sizes, 2, 4, 8, and 16). |
| Duncan et al. (2012), Science | Similar objects are more accurately identified as being similar if they are preceded by new objects than if they are preceded by old objects (a comparison of the fraction of objects rated as similar in trials where they are preceded by new objects compared to trials where they are preceded by old objects in Study 1b (within-subject variation), $t(14) = 3.41$, $p = 0.0042$, p. 486). |
| Gervais and Norenzayan (2012), Science | Priming analytic thinking via images of 'The Thinker' increases religious disbelief compared to viewing control images of a visually similar artwork; a $t$-test, $p < 0.05$ using a two-tailed test. Original test statistics: $N = 57$ (31 in Control condition, 26 in Disbelief condition); Control belief in god (100-pt scale): $M = 61.55$, $SD = 35.68$; Disbelief: $M = 41.42$, $SD = 31.47$; $t(55) = 2.24$; $p = 0.029$ (reported as $p = 0.03$). |
| Gneezy et al. (2014), Science | The likelihood of choosing a charity is higher when potential donors know that the overhead is already paid for, than when the donors pay for overhead themselves (a comparison of the fraction choosing to donate to '*charity: water*' between the '50% overhead, covered treatment' and the '50% overhead treatment', $z = 3.00$, $p < 0.01$ (exact $p = 0.0027$), p. 633). (This hypothesis was picked by lottery instead of comparing the 'no overhead treatment' and the '50% overhead treatment', $z = 3.27$, $p < 0.01$, p. 633.) |

| | |
|---|---|
| Hauser et al. (2014), Nature | Choosing an extraction level for all group members using median voting leads to a higher degree of sustainability of a common pool than allowing each individual to choose their own extraction amount. That is, a comparison of the average probability that the common pool was sustained by the first generation between the voting treatment and the unregulated treatment (in both treatments there is an 80% probability that a new generation occurs and an extraction threshold of 50%). To evaluate this hypothesis, a linear probability model with a treatment dummy variable is used; see the $1^{st}$ generation regression equation in Table S1; $p = 1.427e^{-10}$ (reported as $p < 0.001$) in a $t$-test ($t(38) = 8.696$) of the treatment dummy variable coefficient. |
| Janssen et al. (2010), Science | Communication increases average earnings in a common-pool resource game with spatial and temporal resource dynamics. A comparison of net earnings between the *NCP* condition and the *C* condition in periods 1 to 3 showed $p$-value $< 0.001$ with the Mann-Whitney test ($z = 5.761$ and $p = 8.362e^{-9}$). |
| Karpicke and Blunt (2011), Science | In a memory test one week after learning, Retrieval Practice leads to participants recalling more correct information than Concept-Mapping. A $t$-test, $p < 0.05$ using a two-tailed test, comparing the Retrieval Practice and Concept Mapping conditions. Original test statistics: $N = 40$ (20 in each condition); Mean performance= 0.67 in the Retrieval Practice condition and 0.45 in the Concept Mapping condition. The comparison between Retrieval Practice and Concept Mapping was reported as $F(1, 38) = 21.63$; $p = 0.000039$. |
| Kidd and Castano (2013), Science | Reading literary fiction improves affective Theory of Mind (a comparison of the mean Reading the Mind in the Eyes Test (RMET) score between the literary fiction treatment and the nonfiction treatment in experiment 1; ANOVA test, $F(1, 82) = 6.40$ and $p = 0.0133$ (reported as $p = 0.01$, p. 378). |
| Kovacs et al. (2010), Science | Participants automatically project agents' beliefs and store them in a way similar to that of their own representation about the environment. A comparison of the mean reaction time between the 'P-A- treatment' and the 'P-A+ treatment' in Study 1 (within subject variation), shows that reaction time is shorter in the P-A+ treatment; results show that $t(23) = 2.42$, $p$-value $= 0.02$ (exact $p = 0.0238$). |
| Lee and Schwarz (2010), Science | Hand washing will significantly reduce the need to justify one's choice by increasing the perceived difference between alternatives. Specifically, the mean difference between the rankings of the chosen and rejected albums before and after making the choice will be greater for the soap examining condition compared to the soap hand washing condition. $F$-test assessing the interaction between before-after and hand-washing condition, $p < 0.05$. Original test statistics: (i) *Soap examining condition:* Mean difference between chosen and rejected, before making choice: $M = 0.14$, $SD = 1.01$. Mean difference between chosen and rejected, after making choice: $M = 2.05$, $SD = 1.96$. (ii) *Soap hand washing condition:* Mean difference between chosen and rejected, before making choice: $M = 0.68$, $SD = 0.75$. Mean difference between chosen and rejected, after making choice: $M = 1.00$, $SD = 1.41$. Interaction of before-after and hand-washing: $F(1, 38) = 6.74$, $p = 0.0133$ (reported as $p = 0.01$). |

| | |
|---|---|
| Morewedge et al. (2010), Science | Repeatedly imagining eating a food subsequently reduces the actual consumption of that food (a comparison of the 30-repetition treatment and the control treatment in experiment 1; independent samples $t$-test, $t(30) = 2.78$, $p = 0.0092$, provided by the original authors. The analysis in the original study pools the variance across the 30-repetition, the 3-repetition, and the control condition and reports an ANOVA result of $F(1, 46) = 4.50$, $p = 0.0393$, p. 1531.) (This hypothesis was picked by lottery instead of comparing the mean consumption of M&M's between the 30-repetition treatment and the 3-repetition treatment; $F(1, 46) = 5.81$, $p < 0.05$, p. 1531). |
| Nishi et al. (2015), Nature | In initially unequal situations, wealth visibility leads to greater inequality than when wealth is invisible (a comparison of the mean Gini coefficient between the visible and high initial inequality treatment and the invisible and high initial inequality treatment; OLS regression of the session/round Gini coefficient as the dependent variable and multiway clustering of standard errors at the session and round level; regression equation (5) in Table S2, $p = 0.0044$ of a $t$-test of the treatment dummy variable coefficient, $t(198) = 2.881$). |
| Pyc and Rawson (2010), Science | Retrieval of mediators is greater with test-restudy practice than with restudy practice; a comparison of mean mediator retrieval between the test-restudy and the restudy treatments within the CMR treatment, p. 335, $t(34) = 2.37$ and $p$-value $= 0.02$, $t$-value and $p$-value from authors). Note that a successful retrieval in each of the final test questions is defined as correctly recalling any of the keyword mediators that had been generated during session 1. |
| Ramirez and Beilock (2011), Science | In a high-pressure in-lab math test, those writing for 10 minutes about their deepest thoughts and feelings regarding the upcoming test improve more on that test compared to simply sitting quietly; an $F$-test, $p < 0.05$ using a two-tailed test. Original test statistics: $N = 20$ (10 in each condition); Expressive writing $\mathrm{M}_{pre} = 0.86$ ($SD = 0.09$), $\mathrm{M}_{post} = 0.91$ ($SD = 0.05$), Control $\mathrm{M}_{pre} = 0.82$ ($SD = 0.09$), $\mathrm{M}_{post} = 0.70$ ($SD = 0.11$); $F(1, 18) = 30.53$; $p = 0.00003$ (reported as $p < 0.01$, p. S11). |
| Rand et al. (2012), Nature | Priming intuition increases cooperation in a public goods game compared to priming reflection (a comparison of the mean contribution in a public goods game between the 'intuition-good'/'reflection-bad' treatments and the 'intuition-bad'/'reflection-good' treatments; a Tobit regression (with robust standard errors) with a treatment dummy variable, regression equation (1) in Table S11; $z = 2.617$, $p = 0.0089$ in a $z$-test of the treatment dummy variable coefficient). |
| Shah et al. (2012), Science | Low-wealth subjects, that are given fewer chances to win in repeated 'Wheel of Fortune' type word puzzle games, perform worse in a subsequent attention task (Dots-Mixed task) than do high-wealth individuals (a comparison of the mean performance on the Dots-Mixed task between the 'poor treatment' and the 'rich treatment'; ANOVA test, $F(1, 54) = 4.16$ and $p = 0.046$, p. 683). |
| Sparrow et al. (2011), Science | Computer terms are more accessible than general words after answering a block of hard trivia questions; measured as longer color-naming reaction times in a Modified Stroop Task after priming with computer terms compared to priming with non-computer terms (paired $t$-test, within subject variation; $t(45) = 3.26$, $p = 0.0021$, study 1, p. 776, and Fig. 1). |

| | |
|---|---|
| Wilson et al. (2014), Science | An external activity from a list (e.g. watching television or reading a book) for 12 minutes is rated as being more enjoyable than a 12 minute 'thinking period' entertaining themselves with their thoughts (a higher average self-rated enjoyment (the mean of three nine-point scale items) in the 'external activities' treatment than in the 'standard thought instructions' treatment in Study 8, $t(28) = 4.83$, $p = 0.000044$, p. 76). |

**Supplementary Table 2. Information about whether the original authors shared the materials of their study and approved the replication plan, and if the replication used the same software as the original study.**

| Study | Authors Shared Materials | Replication Used Original Software[†] | Authors Approved Replication Plan |
|---|---|---|---|
| Ackerman et al. (2010), Science | ✓ | ○ | ✓ |
| Aviezer et al. (2012), Science | ✓ | ✗ | ✓ |
| Balafoutas and Sutter (2012), Science | ✓ | ✓ | ✓ |
| Derex et al. (2013), Nature | ✓ | ✓ | ✓ |
| Duncan et al. (2012), Science | ✓ | ✓ | ✓ |
| Gervais and Norenzayan (2012), Science | ✓ | ✓ | ✓ |
| Gneezy et al. (2014), Science | ✓ | ✓ | ✓ |
| Hauser et al. (2014), Nature | ✓ | ✓ | ✓ |
| Janssen et al. (2010), Science | ✓ | ✓ | ✓ |
| Karpicke and Blunt (2011), Science | ✓ | ✗ | ✓ |
| Kidd and Castano (2013), Science | ✓ | ✓ | ✓ |
| Kovacs et al. (2010), Science | ✓ | ✗ | ✓ |
| Lee and Schwarz (2010), Science | ✓ | ○ | ✓ |
| Morewedge et al. (2010), Science | ✓ | ✗ | ✓ |
| Nishi et al. (2015), Nature | ✓ | ✗ | ✓ |
| Pyc and Rawson (2010), Science | ✓ | ✓ | ✓ |
| Ramirez and Beilock (2011), Science | ✓ | ✗ | ✗ |
| Rand et al. (2012), Nature | ✓ | ✗ | ✓ |
| Shah et al. (2012), Science | ✓ | ✓ | ✓ |
| Sparrow et al. (2011), Science | ✗ | ✗ | ✗[*] |
| Wilson et al. (2014), Science | ✓ | ✓ | ✓ |

*Notes:* ✓ indicates "yes", ✗ indicates "no", and ○ denotes "not applicable".
[†] See section 1.2 in the Supplementary Information for details about when the original software was not used.
[*] The original authors did not respond to our requests for materials and feedback on the replication report, prior to conducting the replication.

**Supplementary Table 3. Replication results for Stage 1.**

| Study | Original Study | | | Replication Stage 1 | | | Rep.[†] | Stat. Power[‡] | Relative Effect Size[§] |
|---|---|---|---|---|---|---|---|---|---|
| | Effect Size ($r$) | $p$-Value | $N^*$ | Effect Size ($r$) | $p$-Value | $N^*$ | | | |
| Ackerman et al. (2010), Science | 0.270 | 0.049 | 54 (54) | 0.141 | 0.024 | 259 (259) | yes | 0.901 | 0.521 |
| Aviezer et al. (2012), Science | 0.961 | $< 0.001$ | 15 (15) | 0.829 | $< 0.001$ | 14 (14) | yes | 0.930 | 0.862 |
| Balafoutas and Sutter (2012), Science | 0.278 | 0.018 | 72 (72) | 0.146 | 0.022 | 243 (243) | yes | 0.898 | 0.527 |
| Derex et al. (2013), Nature | 0.525 | $< 0.001$ | 51 (366) | 0.361 | 0.003 | 65 (482) | yes | 0.902 | 0.687 |
| Duncan et al. (2012), Science | 0.674 | 0.004 | 15 (15) | 0.183 | 0.279 | 36 (36) | no | 0.909 | 0.271 |
| Gervais and Norenzayan (2012), Science | 0.289 | 0.029 | 57 (57) | $-0.055$ | 0.416 | 224 (224) | no | 0.902 | $-0.190$ |
| Gneezy et al. (2014), Science | 0.223 | 0.003 | 178 (178) | 0.182 | $< 0.001$ | 407 (407) | yes | 0.922 | 0.818 |
| Hauser et al. (2014), Nature | 0.816 | $< 0.001$ | 40 (200) | 0.832 | $< 0.001$ | 22 (110) | yes | 0.919 | 1.020 |
| Janssen et al. (2010), Science | 0.631 | $< 0.001$ | 63 (105) | 0.344 | 0.025 | 42 (70) | yes | 0.902 | 0.545 |
| Karpicke and Blunt (2011), Science | 0.602 | $< 0.001$ | 40 (40) | 0.384 | 0.006 | 49 (49) | yes | 0.922 | 0.638 |
| Kidd and Castano (2013), Science | 0.269 | 0.013 | 86 (86) | $-0.066$ | 0.273 | 285 (285) | no | 0.923 | $-0.244$ |
| Kovacs et al. (2010), Science | 0.450 | 0.024 | 24 (24) | 0.586 | $< 0.001$ | 95 (95) | yes | 0.923 | 1.301 |
| Lee and Schwarz (2010), Science | 0.388 | 0.013 | 40 (40) | $-0.068$ | 0.455 | 123 (123) | no | 0.904 | $-0.176$ |
| Morewedge et al. (2010), Science | 0.453 | 0.009 | 32 (32) | 0.355 | $< 0.001$ | 89 (89) | yes | 0.904 | 0.783 |
| Nishi et al. (2015), Nature | 0.201 | 0.004 | 200 (366) | 0.116 | 0.011 | 480 (792) | yes | 0.912 | 0.579 |
| Pyc and Rawson (2010), Science | 0.377 | 0.024 | 36 (36) | 0.148 | 0.089 | 132 (132) | no | 0.904 | 0.394 |
| Ramirez and Beilock (2011), Science | 0.793 | $< 0.001$ | 20 (20) | $-0.075$ | 0.716 | 26 (52) | no | 0.929 | $-0.095$ |
| Rand et al. (2012), Nature | 0.141 | 0.009 | 343 (343) | 0.028 | 0.366 | 1014 (1014) | no | 0.920 | 0.202 |
| Shah et al. (2012), Science | 0.267 | 0.046 | 56 (56) | $-0.087$ | 0.150 | 278 (278) | no | 0.916 | $-0.326$ |
| Sparrow et al. (2011), Science | 0.368 | 0.002 | 69 (69) | 0.110 | 0.265 | 104 (104) | no | 0.820 | 0.299 |
| Wilson et al. (2014), Science | 0.674 | $< 0.001$ | 30 (30) | 0.594 | $< 0.001$ | 39 (39) | yes | 0.930 | 0.880 |

* Number of observations; number of individuals provided in parenthesis.

† Replicated; significant effect ($p < 0.05$) in the same direction as in original study.

‡ Statistical power to detect 75% of the original effect size $r$.

§ Relative standardized effect size.

**Supplementary Table 4. Replication results for Stage 2.**

| Study | Original Study | | | Replication Stage 2 | | | Rep.[†] | Stat. Power[‡] | Relative Effect Size[§] |
|---|---|---|---|---|---|---|---|---|---|
| | Effect Size ($r$) | $p$-Value | $N^*$ | Effect Size ($r$) | $p$-Value | $N^*$ | | | |
| Ackerman et al. (2010), Science | 0.270 | 0.049 | 54 (54) | 0.063 | 0.125 | 599 (599) | no | 0.904 | 0.232 |
| Aviezer et al. (2012), Science | 0.961 | $< 0.001$ | 15 (15) | | | | | | |
| Balafoutas and Sutter (2012), Science | 0.278 | 0.018 | 72 (72) | | | | | | |
| Derex et al. (2013), Nature | 0.525 | $< 0.001$ | 51 (366) | | | | | | |
| Duncan et al. (2012), Science | 0.674 | 0.004 | 15 (15) | 0.436 | $< 0.001$ | 92 (92) | yes | 0.906 | 0.648 |
| Gervais and Norenzayan (2012), Science | 0.289 | 0.029 | 57 (57) | $-0.035$ | 0.415 | 531 (531) | no | 0.910 | $-0.123$ |
| Gneezy et al. (2014), Science | 0.223 | 0.003 | 178 (178) | | | | | | |
| Hauser et al. (2014), Nature | 0.816 | $< 0.001$ | 40 (200) | | | | | | |
| Janssen et al. (2010), Science | 0.631 | $< 0.001$ | 63 (105) | | | | | | |
| Karpicke and Blunt (2011), Science | 0.602 | $< 0.001$ | 40 (40) | | | | | | |
| Kidd and Castano (2013), Science | 0.269 | 0.013 | 86 (86) | $-0.027$ | 0.468 | 714 (714) | no | 0.943 | $-0.101$ |
| Kovacs et al. (2010), Science | 0.450 | 0.024 | 24 (24) | | | | | | |
| Lee and Schwarz (2010), Science | 0.388 | 0.013 | 40 (40) | $-0.046$ | 0.436 | 286 (286) | no | 0.901 | $-0.119$ |
| Morewedge et al. (2010), Science | 0.453 | 0.009 | 32 (32) | | | | | | |
| Nishi et al. (2015), Nature | 0.201 | 0.004 | 200 (366) | | | | | | |
| Pyc and Rawson (2010), Science | 0.377 | 0.024 | 36 (36) | 0.150 | 0.009 | 306 (306) | yes | 0.901 | 0.398 |
| Ramirez and Beilock (2011), Science | 0.793 | $< 0.001$ | 20 (20) | $-0.098$ | 0.394 | 79 (131) | no | 0.949 | $-0.124$ |
| Rand et al. (2012), Nature | 0.141 | 0.009 | 343 (343) | 0.026 | 0.234 | 2136 (2136) | no | 0.901 | 0.183 |
| Shah et al. (2012), Science | 0.267 | 0.046 | 56 (56) | $-0.015$ | 0.710 | 619 (619) | no | 0.908 | $-0.056$ |
| Sparrow et al. (2011), Science | 0.368 | 0.002 | 69 (69) | 0.050 | 0.449 | 234 (234) | no | 0.807 | 0.135 |
| Wilson et al. (2014), Science | 0.674 | $< 0.001$ | 30 (30) | | | | | | |

* Number of observations; number of individuals provided in parenthesis.

† Replicated; significant effect ($p < 0.05$) in the same direction as in original study.

‡ Statistical power to detect 50% of the original effect size $r$.

§ Relative standardized effect size.

**Supplementary Table 5. Prediction market and survey beliefs for the 21 replication studies in Treatment 1 and Treatment 2.**

| | Treatment 1 | | Treatment 2 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Study** | *Market Belief s1** | *Survey Belief s1** | *Market Belief s1** | *Survey Belief s1** | *Market Belief s2†* | *Survey Belief s2†* | *Market Belief s1+s2* | *Survey Belief s1+s2* | *Rep. s1* | *Rep. s1+s2* |
| Ackerman et al. (2010), Science | 0.23 | 0.19 | 0.15 | 0.13 | 0.08 | 0.15 | 0.23 | 0.28 | yes | no |
| Aviezer et al. (2012), Science | 0.66 | 0.50 | 0.49 | 0.43 | 0.31 | 0.25 | 0.80 | 0.68 | yes | |
| Balafoutas and Sutter (2012), Science | 0.75 | 0.56 | 0.75 | 0.43 | 0.13 | 0.27 | 0.88 | 0.70 | yes | |
| Derex et al. (2013), Nature | 0.63 | 0.65 | 0.51 | 0.50 | 0.14 | 0.27 | 0.65 | 0.77 | yes | |
| Duncan et al. (2012), Science | 0.72 | 0.50 | 0.56 | 0.39 | 0.18 | 0.27 | 0.74 | 0.65 | no | yes |
| Gervais and Norenzayan (2012), Science | 0.21 | 0.29 | 0.17 | 0.20 | 0.21 | 0.18 | 0.38 | 0.38 | no | no |
| Gneezy et al. (2014), Science | 0.78 | 0.71 | 0.83 | 0.54 | 0.10 | 0.26 | 0.94 | 0.80 | yes | |
| Hauser et al. (2014), Nature | 0.80 | 0.70 | 0.83 | 0.52 | 0.13 | 0.27 | 0.96 | 0.80 | yes | |
| Janssen et al. (2010), Science | 0.79 | 0.68 | 0.69 | 0.54 | 0.21 | 0.28 | 0.90 | 0.82 | yes | |
| Karpicke and Blunt (2011), Science | 0.73 | 0.64 | 0.49 | 0.48 | 0.23 | 0.30 | 0.72 | 0.78 | yes | |
| Kidd and Castano (2013), Science | 0.39 | 0.37 | 0.28 | 0.22 | 0.06 | 0.24 | 0.34 | 0.46 | no | no |
| Kovacs et al. (2010), Science | 0.47 | 0.40 | 0.39 | 0.29 | 0.23 | 0.24 | 0.63 | 0.53 | yes | |
| Lee and Schwarz (2010), Science | 0.23 | 0.23 | 0.24 | 0.15 | 0.09 | 0.17 | 0.33 | 0.32 | no | no |
| Morewedge et al. (2010), Science | 0.50 | 0.41 | 0.28 | 0.30 | 0.31 | 0.24 | 0.59 | 0.55 | yes | |
| Nishi et al. (2015), Nature | 0.71 | 0.61 | 0.56 | 0.49 | 0.22 | 0.26 | 0.78 | 0.75 | yes | |
| Pyc and Rawson (2010), Science | 0.74 | 0.45 | 0.58 | 0.34 | 0.23 | 0.26 | 0.82 | 0.60 | no | yes |
| Ramirez and Beilock (2011), Science | 0.56 | 0.42 | 0.26 | 0.31 | 0.26 | 0.24 | 0.52 | 0.54 | no | no |
| Rand et al. (2012), Nature | 0.40 | 0.51 | 0.34 | 0.33 | 0.19 | 0.22 | 0.53 | 0.55 | no | no |
| Shah et al. (2012), Science | 0.38 | 0.36 | 0.28 | 0.23 | 0.20 | 0.19 | 0.49 | 0.41 | no | no |
| Sparrow et al. (2011), Science | 0.51 | 0.44 | 0.40 | 0.33 | 0.11 | 0.24 | 0.51 | 0.57 | no | no |
| Wilson et al. (2014), Science | 0.75 | 0.65 | 0.46 | 0.52 | 0.11 | 0.26 | 0.57 | 0.78 | yes | |

* Belief about the probability of replicating in stage 1 (90% power to detect 75% of the original effect size).

† Predicted added probability of replicating in stage 2 (90% power to detect 50% of the original effect size) compared to stage 1.

**Supplementary Table 6. Additional prediction market data for the 21 replication studies.**

| Study | Treatment 1 | | | | Treatment 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | *Tokens Invested** | *Volume (Shares)*† | *Trans-actions* | *No. of Traders* | *Tokens Invested** | *Volume (Shares)*† | *Trans-actions* | *No. of Traders* |
| Ackerman et al. (2010), Science | 9.04 | 17.55 | 244 | 80 | 14.02 | 14.43 | 137 | 68 |
| Aviezer et al. (2012), Science | 6.64 | 13.37 | 75 | 37 | 7.16 | 7.21 | 72 | 36 |
| Balafoutas and Sutter (2012), Science | 9.43 | 17.66 | 70 | 45 | 6.06 | 5.43 | 64 | 35 |
| Derex et al. (2013), Nature | 8.26 | 15.95 | 76 | 47 | 7.42 | 6.28 | 70 | 32 |
| Duncan et al. (2012), Science | 7.22 | 14.03 | 49 | 39 | 5.61 | 5.45 | 55 | 34 |
| Gervais and Norenzayan (2012), Science | 8.17 | 15.34 | 186 | 71 | 8.98 | 8.30 | 106 | 64 |
| Gneezy et al. (2014), Science | 8.50 | 15.84 | 102 | 62 | 6.49 | 5.77 | 89 | 46 |
| Hauser et al. (2014), Nature | 9.54 | 18.18 | 84 | 56 | 8.67 | 8.34 | 77 | 37 |
| Janssen et al. (2010), Science | 8.44 | 15.48 | 55 | 39 | 4.99 | 4.29 | 47 | 27 |
| Karpicke and Blunt (2011), Science | 5.79 | 11.00 | 65 | 39 | 7.65 | 7.19 | 81 | 41 |
| Kidd and Castano (2013), Science | 6.64 | 12.52 | 162 | 51 | 7.32 | 7.11 | 109 | 57 |
| Kovacs et al. (2010), Science | 4.99 | 9.85 | 69 | 37 | 4.74 | 4.82 | 47 | 25 |
| Lee and Schwarz (2010), Science | 8.17 | 14.80 | 204 | 65 | 8.23 | 7.78 | 90 | 59 |
| Morewedge et al. (2010), Science | 6.20 | 12.37 | 92 | 57 | 8.23 | 7.83 | 71 | 40 |
| Nishi et al. (2015), Nature | 12.18 | 23.85 | 51 | 39 | 6.31 | 6.46 | 60 | 37 |
| Pyc and Rawson (2010), Science | 7.41 | 14.36 | 53 | 36 | 4.49 | 4.05 | 26 | 18 |
| Ramirez and Beilock (2011), Science | 7.05 | 14.37 | 94 | 53 | 8.63 | 9.03 | 103 | 52 |
| Rand et al. (2012), Nature | 6.38 | 12.80 | 79 | 48 | 7.82 | 8.02 | 139 | 52 |
| Shah et al. (2012), Science | 8.09 | 15.44 | 125 | 59 | 8.77 | 8.18 | 61 | 36 |
| Sparrow et al. (2011), Science | 7.11 | 14.00 | 71 | 40 | 6.86 | 6.86 | 72 | 43 |
| Wilson et al. (2014), Science | 10.63 | 20.50 | 103 | 66 | 7.01 | 6.64 | 120 | 53 |

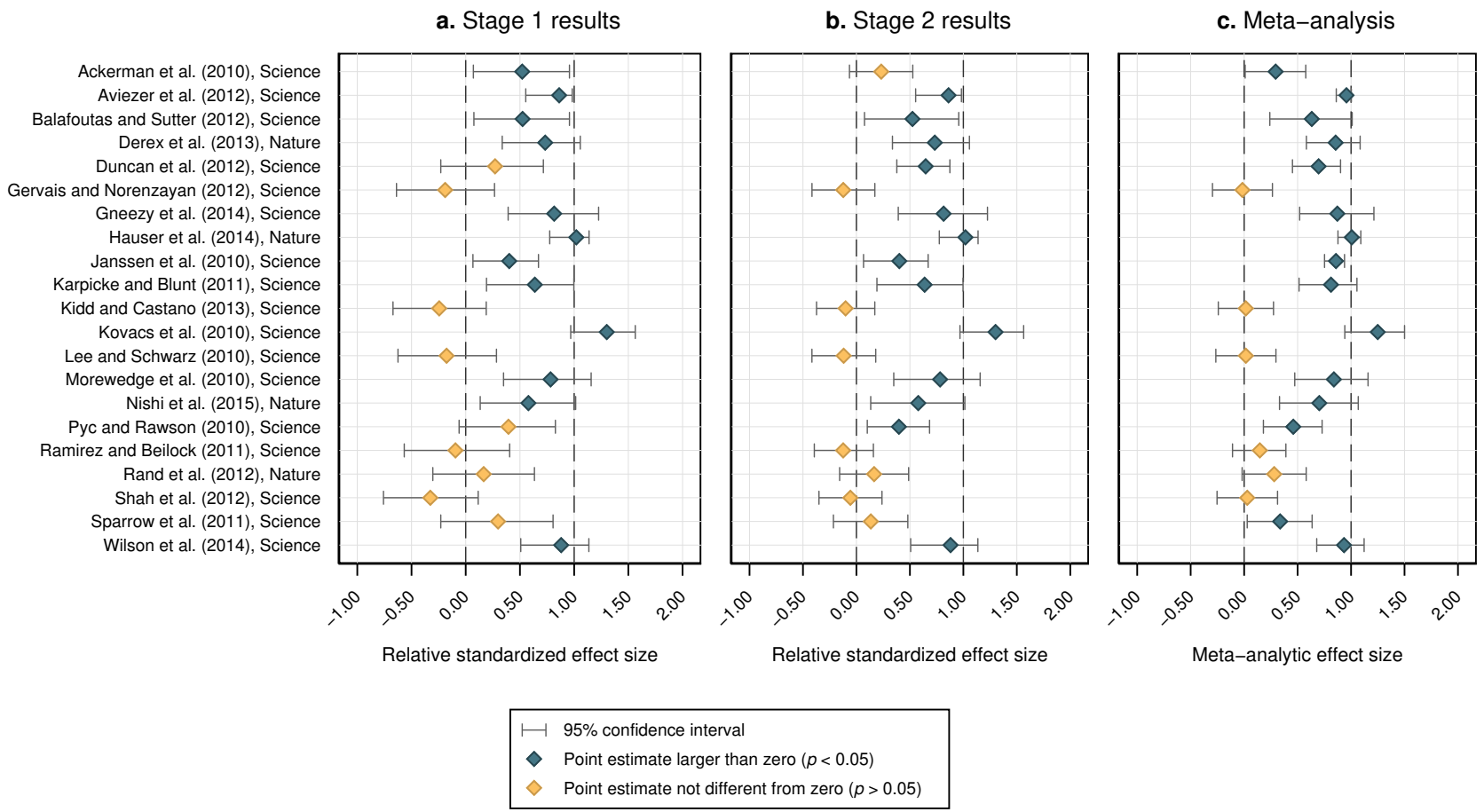
* Mean number of tokens (points) invested per transaction.

† Mean number of shares bought or sold per transaction.

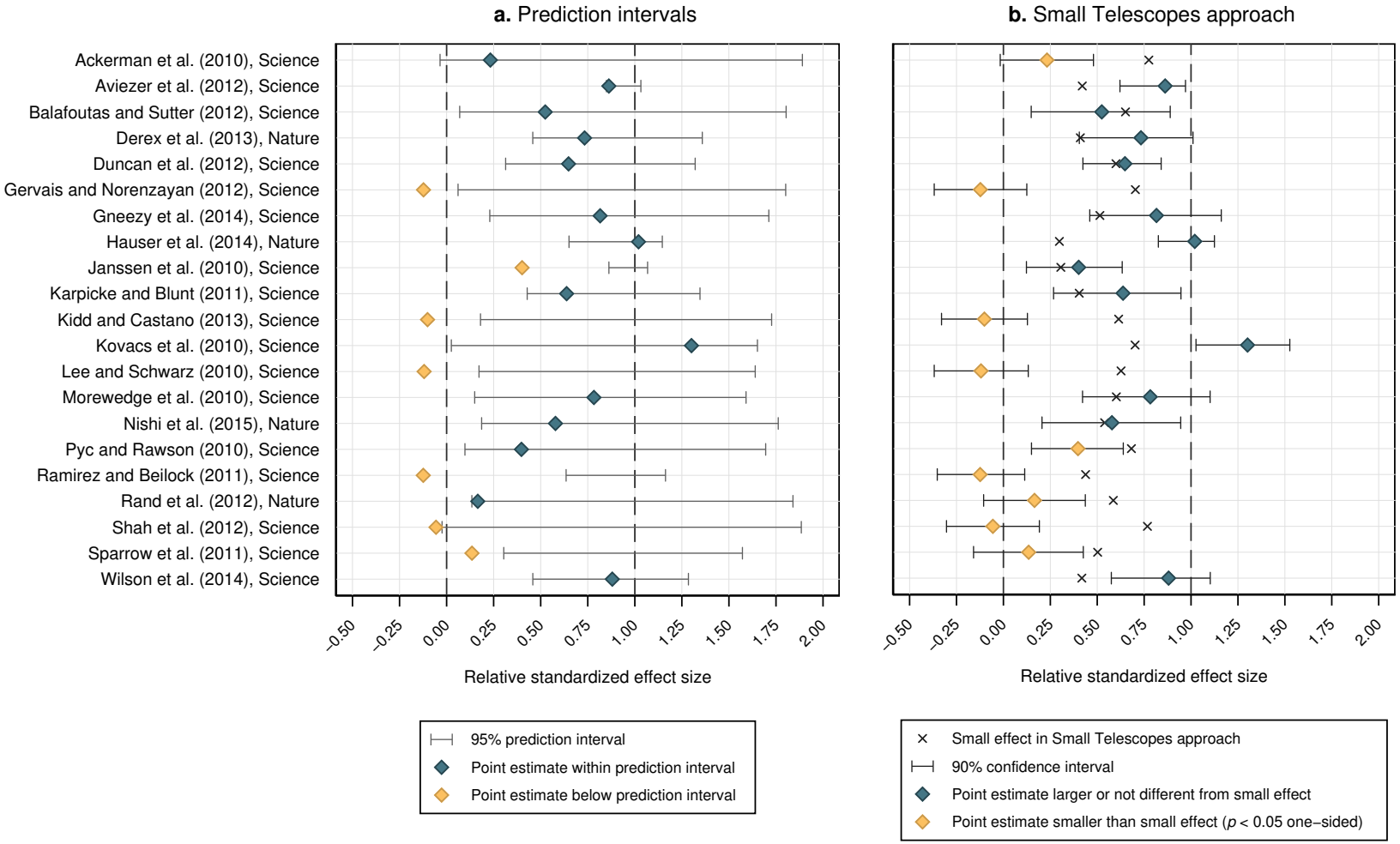
**Supplementary Table 7. Correlation matrix for the reproducibility indicators and the two original study characteristics.** Spearman correlations ($p$-values).

| ***Variable*** | Rep. $p < 0.05$ | Meta Estimate $p < 0.05$ | Small Tele-scopes | Rep. within 95% PI | Relative Effect Size | Default Bayes Factor | Rep. Bayes Factor | Market Belief (T1) | Market Belief (T2) | Survey Belief (T1) | Survey Belief (T2) | Original $p$-Value | Original No. of Obs. | Original No. of Part. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Replicated $p < 0.05$ | 1.0000 | | | | | | | | | | | | | |
| | (0.0000) | | | | | | | | | | | | | |
| Meta Estimate $p < 0.05$ | 0.7126 | 1.0000 | | | | | | | | | | | | |
| | (0.0003) | (0.0000) | | | | | | | | | | | | |
| Small Telescopes Approach | 0.9058 | 0.6455 | 1.0000 | | | | | | | | | | | |
| | (0.0000) | (0.0016) | (0.0000) | | | | | | | | | | | |
| Replication within 95% PI | 0.9014 | 0.7906 | 0.8165 | 1.0000 | | | | | | | | | | |
| | (0.0000) | (0.0000) | (0.0000) | (0.0000) | | | | | | | | | | |
| Relative Effect Size | 0.8421 | 0.7385 | 0.8581 | 0.8174 | 1.0000 | | | | | | | | | |
| | (0.0000) | (0.0001) | (0.0000) | (0.0000) | (0.0000) | | | | | | | | | |
| Default Bayes Factor | 0.8423 | 0.7296 | 0.8266 | 0.8177 | 0.9341 | 1.0000 | | | | | | | | |
| | (0.0000) | (0.0002) | (0.0000) | (0.0000) | (0.0000) | (0.0000) | | | | | | | | |
| Replication Bayes Factor | 0.7213 | 0.6284 | 0.7237 | 0.7097 | 0.8924 | 0.9054 | 1.0000 | | | | | | | |
| | (0.0002) | (0.0023) | (0.0002) | (0.0003) | (0.0000) | (0.0000) | (0.0000) | | | | | | | |
| Market Beliefs (Treatment 1) | 0.7773 | 0.6093 | 0.6833 | 0.6673 | 0.5961 | 0.5898 | 0.4036 | 1.0000 | | | | | | |
| | (0.0000) | (0.0034) | (0.0006) | (0.0010) | (0.0043) | (0.0049) | (0.0696) | (0.0000) | | | | | | |
| Market Beliefs (Treatment 2) | 0.8421 | 0.6278 | 0.7310 | 0.7006 | 0.6416 | 0.6294 | 0.4855 | 0.8896 | 1.0000 | | | | | |
| | (0.0000) | (0.0023) | (0.0002) | (0.0004) | (0.0017) | (0.0022) | (0.0257) | (0.0000) | (0.0000) | | | | | |
| Survey Beliefs (Treatment 1) | 0.6963 | 0.6093 | 0.6992 | 0.5505 | 0.5987 | 0.5788 | 0.4355 | 0.8935 | 0.8169 | 1.0000 | | | | |
| | (0.0005) | (0.0034) | (0.0004) | (0.0097) | (0.0041) | (0.0060) | (0.0485) | (0.0000) | (0.0000) | (0.0000) | | | | |
| Survey Beliefs (Treatment 2) | 0.7611 | 0.6278 | 0.7469 | 0.6172 | 0.6208 | 0.6015 | 0.4225 | 0.9312 | 0.8455 | 0.9805 | 1.0000 | | | |
| | (0.0001) | (0.0023) | (0.0001) | (0.0029) | (0.0027) | (0.0039) | (0.0564) | (0.0000) | (0.0000) | (0.0000) | (0.0000) | | | |
| Original $p$-Value | -0.4048 | -0.3139 | -0.4926 | -0.2502 | -0.3909 | -0.4300 | -0.1722 | -0.6519 | -0.5156 | -0.7247 | -0.7351 | 1.0000 | | |
| | (0.0687) | (0.1659) | (0.0233) | (0.2739) | (0.0797) | (0.0517) | (0.4553) | (0.0014) | (0.0167) | (0.0002) | (0.0001) | (0.0000) | | |
| Original No. of Observations | -0.2920 | -0.0370 | -0.2228 | -0.2841 | -0.3395 | -0.4723 | -0.3561 | -0.0767 | -0.0377 | 0.1398 | 0.0683 | 0.2237 | 1.0000 | |
| | (0.1991) | (0.8735) | (0.3316) | (0.2121) | (0.1321) | (0.0306) | (0.1131) | (0.7409) | (0.8710) | (0.5455) | (0.7687) | (0.3296) | (0.0000) | |
| Original No. of Participants | -0.0567 | 0.1016 | 0.0080 | -0.0751 | -0.0877 | -0.1463 | -0.0843 | 0.1215 | 0.1768 | 0.3880 | 0.3035 | -0.0390 | 0.8568 | 1.0000 |
| | (0.8070) | (0.6611) | (0.9727) | (0.7462) | (0.7053) | (0.5269) | (0.7165) | (0.5997) | (0.4433) | (0.0822) | (0.1810) | (0.8667) | (0.0000) | (0.0000) |

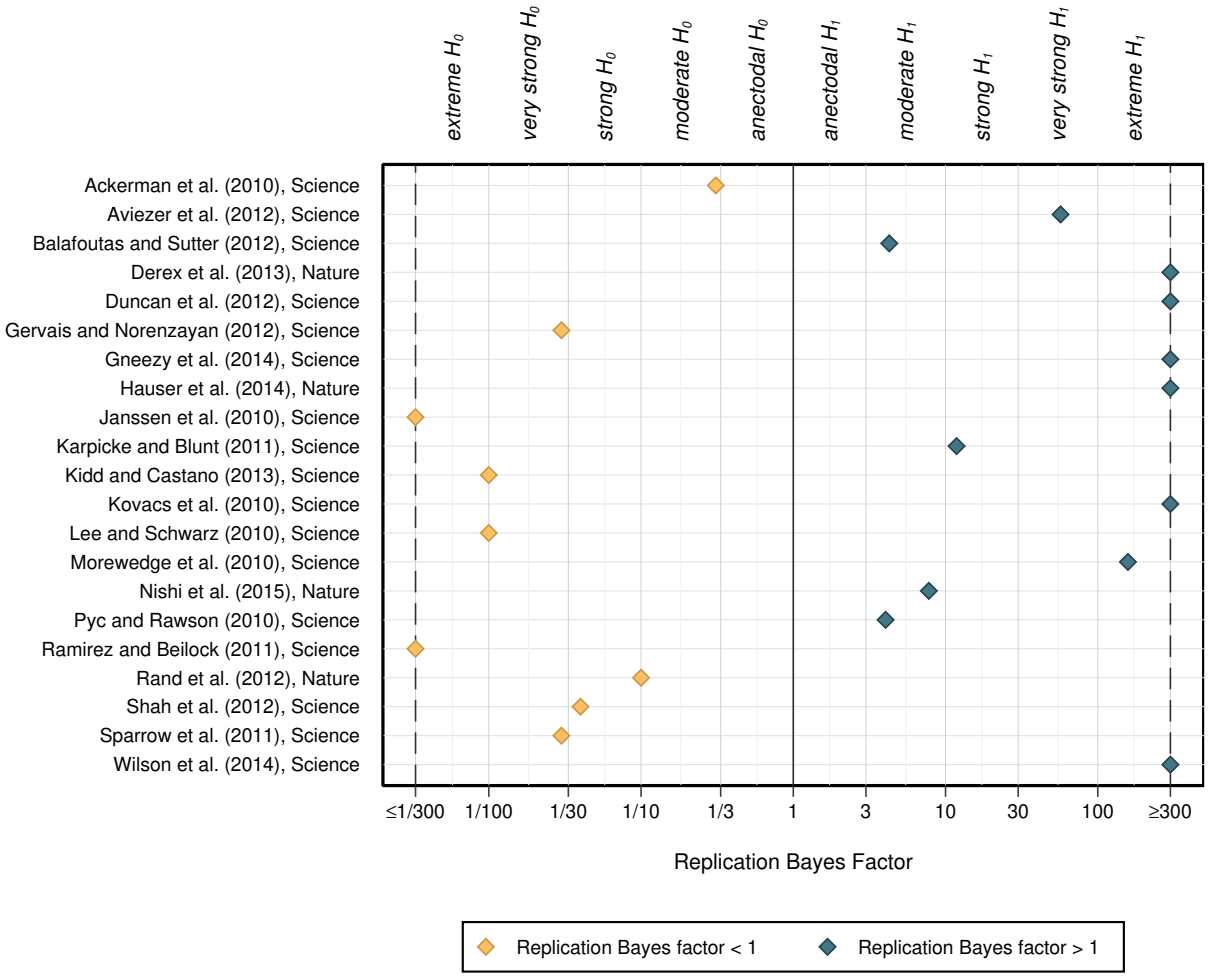
# Supplementary Figures



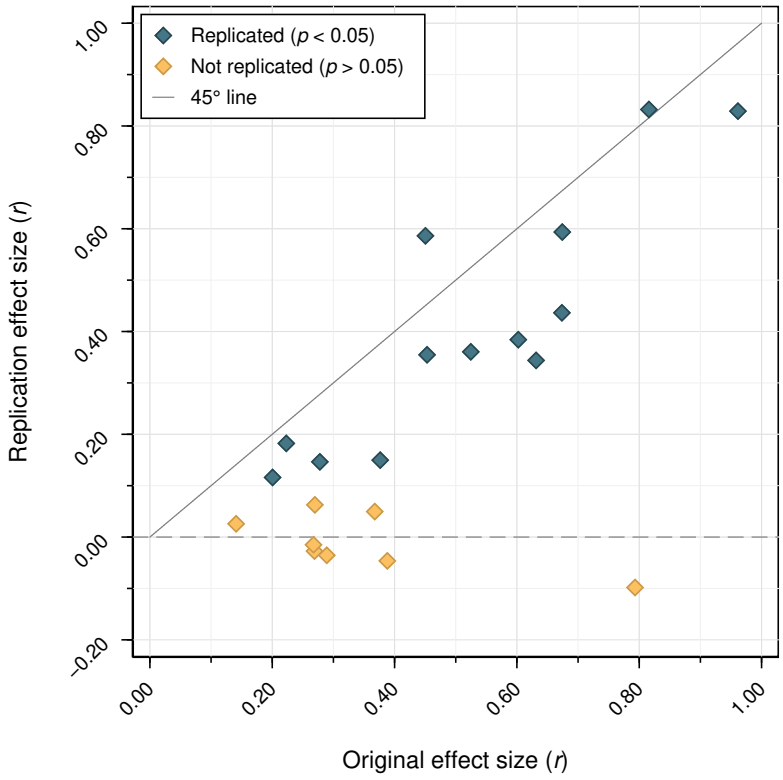

**Supplementary Figure 1. Replication Results after Stage 1 and Stage 2. Robustness test with all results based on *t*-tests or *F*(1, df) tests.** (**a**) Plotted are 95% CIs of replication effect sizes (standardized to correlation coefficients $r$) after Stage 1. The standardized effect sizes are normalized 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 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 replication studies. 95% CIs of standardized effect sizes (correlation coefficient $r$). The standardized effect sizes are normalized so that 1 equals the original effect size. 15 out of 21 studies have a significant effect in the same direction as the original study in the meta-analysis [71.43%; 95% CI = (50.4%, 92.5%)].



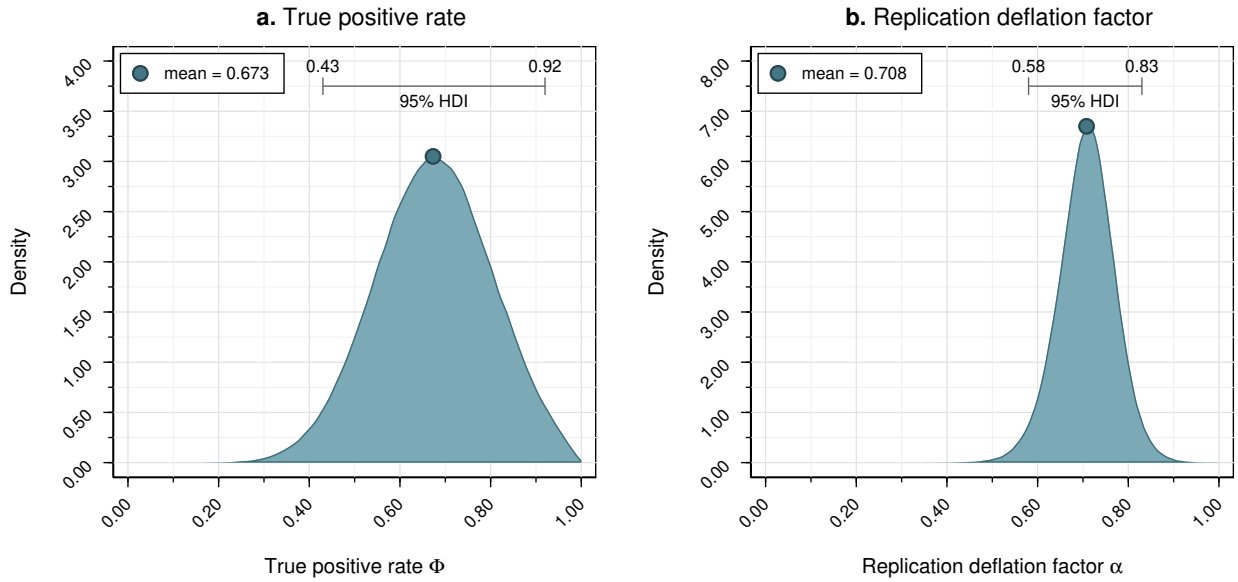

**Supplementary Figure 2. Replication results for two complementary replication indicators; 95% prediction intervals[29] in panel a and the Small Telescopes approach in panel b[30]. Robustness test with all results based on *t*-tests or *F*(1, df) tests.** (**a**) Plotted are 95% prediction intervals for the standardized original effect sizes (correlation coefficient $r$). The standardized effect sizes are normalized so that 1 equals the original effect size. 14 out of 21 replications [66.7%; 95% CI = (44.7%, 88.6%)] are within the 95% prediction interval and replicate according to this indicator. (**b**) Plotted are 90% CIs of replication effect sizes in relation to small effect sizes as defined by the Small Telescopes approach (the effect size 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% confidence interval 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.




**Supplementary Figure 3. Replication Bayes factors for the 21 replications[34].** A replication Bayes factor above one favors the effect size observed in the original study and a replication factor below 1 favors the null hypothesis of no effect. The evidence categories proposed by Jeffreys[33] are also shown in the figure (from extreme support for the null hypothesis of no effect to extreme support for the original study effect size).




**Supplementary Figure 4. Original study effect size versus replication effect size (correlation coefficients $r$).** The diagonal line represents a replication effect size equal to the original effect size and the dotted line represents a replication effect size equal to zero. The mean standardized effect size (correlation coefficient, $r$) of the replications is 0.249 ($SD = 0.283$), compared to 0.459 ($SD = 0.229$) in the original studies. This difference is significant (Wilcoxon signed-ranks test, $n = 21$, $z = 3.667$, $p < 0.001$). The mean *relative* effect size of all the replications is 46.2% [95% CI = (27.0%, 65.5%)]; the mean *relative* effect size of the replications that replicated is 74.5% [95% CI = (60.1%, 88.9%)]; and the mean relative effect size of the replications that did not replicate is 0.3% [95% CI = (−12.4%, 13.1%)]. The Spearman correlation between the original effect size and the replication effect size is 0.574 [$p = 0.007$; 95% CI = (18.9%, 80.6%)].

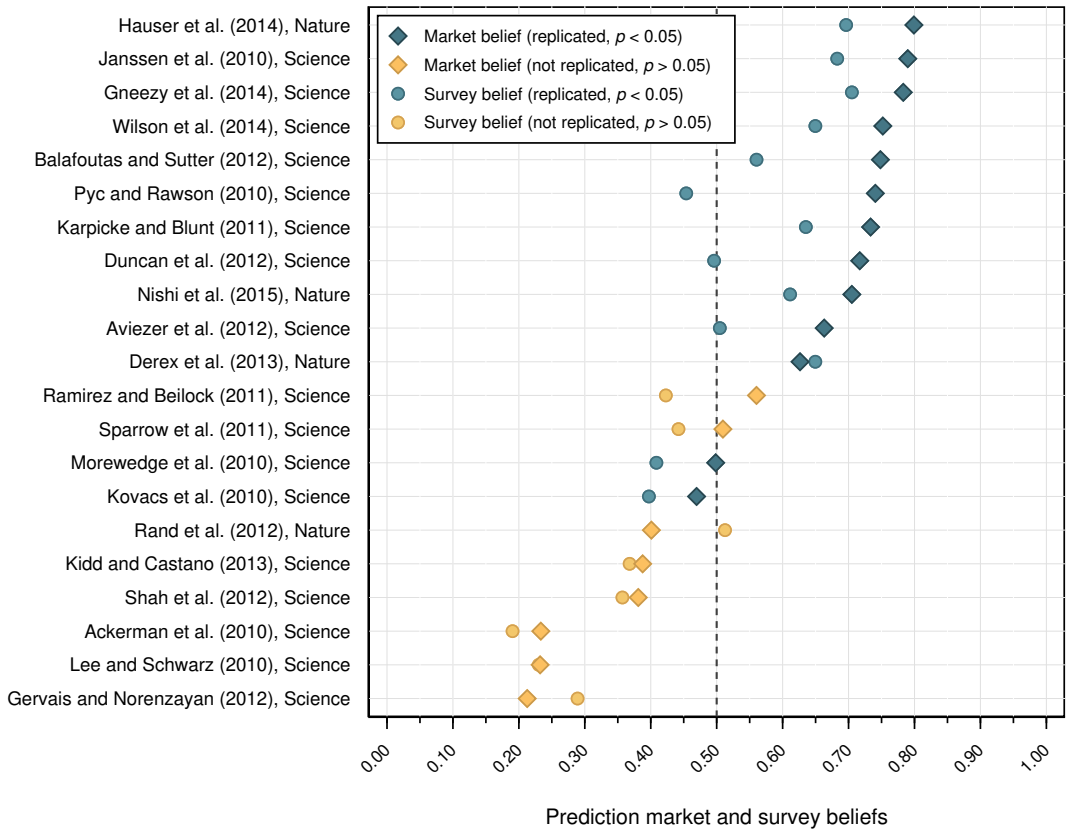



**Supplementary Figure 5. Bayesian inference for the errors-in-variables mixture model[35–37].** (**A**) Posterior distribution of the true positive rate. The posterior mean is 0.673 with a 95% credible interval that ranges from 0.43 to 0.92. (**b**) Posterior distribution of the replication deflation factor, that is the degree to which true effects are overestimated in original studies vs. replication studies. The posterior mean is 0.708, with a 95% credible interval that ranges from 0.58 to 0.83.

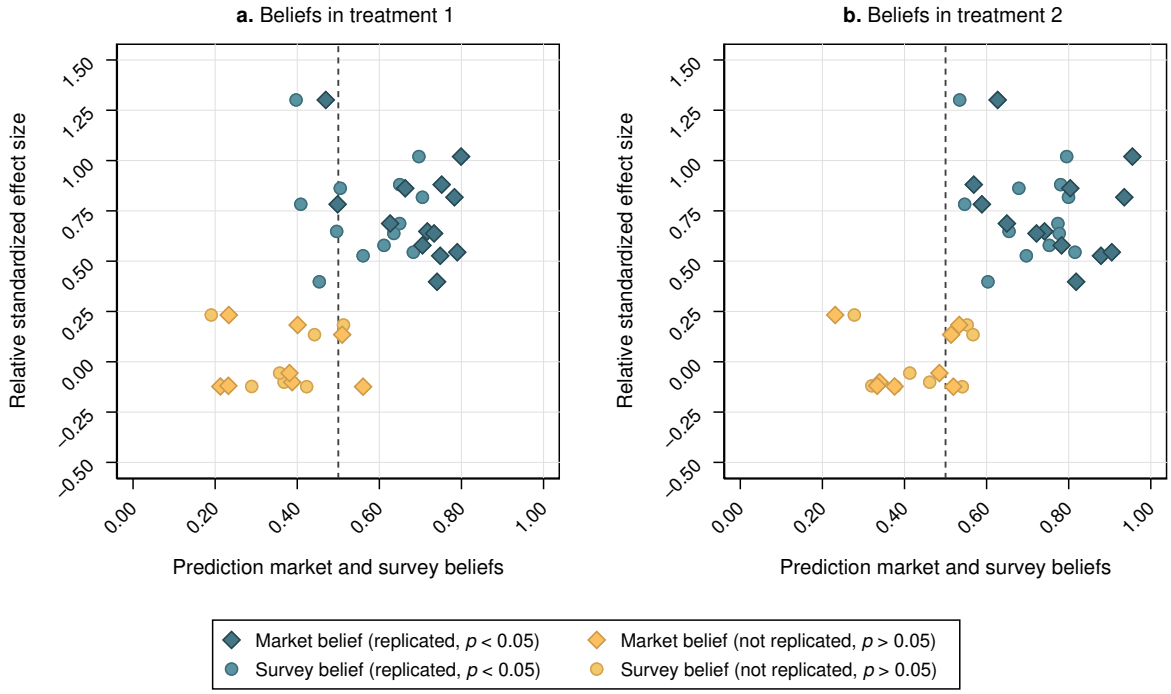
**Panel A**


**Panel B**

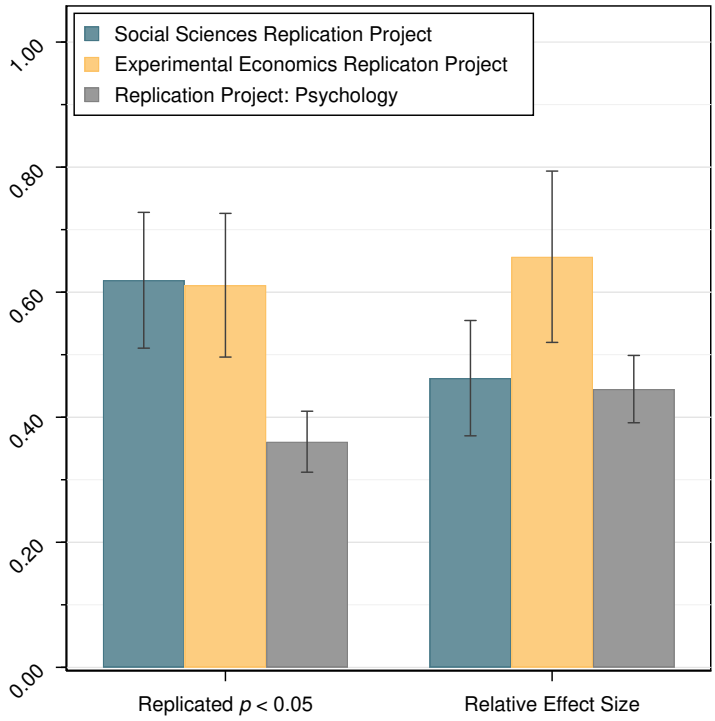

**Supplementary Figure 6. Trading interface.** (**a**) Trading interface for Treatment 1. (**b**) Trading interface for Treatment 2.




**Supplementary Figure 7. Prediction market and survey beliefs.** The figure shows the prediction market beliefs and the survey beliefs of replicating after Stage 1 (from Treatment 1 for measuring beliefs). The replication studies are ranked in terms of prediction market beliefs on the $y$-axis. The mean prediction market belief for replicating after Stage 1 is 56.9% [range 21.3% to 79.9%, 95% CI = (47.8%, 65.9%)], and the mean survey belief is 48.9% [range 19.0% to 70.5%, 95% CI = (41.9%, 55.9%)]. The prediction market beliefs and survey beliefs are highly correlated [Spearman correlation coefficient 0.894, $p < 0.001$, 95% CI = $(0.752, 0.956)$, $n = 21$]. Both the prediction market beliefs and the survey beliefs are also highly correlated with a successful replication after Stage 1 [Spearman correlation coefficient for prediction market beliefs 0.509, $p = 0.019$, 95% CI = (9.8%, 77.1%), $n = 21$; Spearman correlation coefficient for survey beliefs 0.540, $p = 0.011$, 95% CI = (14.2%, 78.8%), $n = 21$] and a successful replication after Stage 2 [Spearman correlation coefficient for prediction market beliefs 0.777, $p < 0.001$, 95% CI = (52.0%, 90.5%), $n = 21$; Spearman correlation coefficient for survey beliefs 0.696, $p =< 0.001$, 95% CI = (37.8%, 86.7%), $n = 21$].




**Supplementary Figure 8. Prediction market and survey beliefs and the relative effect size.** (**a**) Plotted are prediction market and survey beliefs about replicating after Stage 1 from Treatment 1 and relative effect sizes of the replications (with relative effect sizes based on the maximum data after Stage 2). Both the prediction market beliefs [Spearman correlation coefficient 0.596, $p = 0.004$, 95% CI = (22.1%, 81.7%), $n = 21$], and the survey beliefs [Spearman correlation coefficient 0.599, $p = 0.004$, 95% CI = (22.5%, 81.9%), $n = 21$] are positively correlated with the relative effect size of the replications. (**b**) Plotted are prediction market and survey beliefs about replicating after Stage 2 from Treatment 2 and relative effect sizes of the replications. Both the prediction market beliefs [Spearman correlation coefficient 0.642, $p = 0.002$, 95% CI = (29.0%, 84.0%), $n = 21$], and the survey beliefs [Spearman correlation coefficient 0.621, $p = 0.003$, 95% CI = (25.8%, 83.0%), $n = 21$] are positively correlated with the relative effect size of the replications.




**Supplementary Figure 9. A comparison of replicability indicators between the Social Sciences Replication Project (SSRP), the Experimental Economics Replication Project (EERP)[8], and the Reproducibility Project: Psychology (RPP)[7].** Error bars denotes ±se.