Replication by plebiscite: Examining the replicability of online experiments selected by a decision market

Felix Holzmeister[1], Magnus Johannesson[2], Colin F. Camerer[3], Yiling Chen[4], Teck-Hua Ho[5], Suzanne Hoogeveen[6], Juergen Huber[7], Noriko Imai[8], Taisuke Imai[8], Lawrence Jin[9], Michael Kirchler[7], Alexander Ly[10,11], Benjamin Mandl[12], Dylan Manfredi[13], Gideon Nave[13], Brian A. Nosek[14], Thomas Pfeiffer[15], Alexandra Sarafoglou[10], Rene Schwaiger[7], Eric-Jan Wagenmakers[10], Viking Waldén[2], Anna Dreber[1,2,*]

[1] Department of Economics, University of Innsbruck, Innsbruck, Austria. [2] Department of Economics, Stockholm School of Economics, Stockholm, Sweden. [3] Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, USA. [4] John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, USA. [5] Nanyang Technological University, Singapore. [6] Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands. [7] Department of Banking and Finance, University of Innsbruck, Innsbruck, Austria. [8] Institute of Social and Economic Research, Osaka University, Osaka, Japan. [9] Lee Kuan Yew School of Public Policy, National University of Singapore, Singapore. [10] Faculty of Social and Behavioural Sciences, University of Amsterdam, Amsterdam, [11] Machine Learning, Centrum Wiskunde and Informatica, Amsterdam, The Netherlands. [12] enspired, Vienna, Austria. [13] Marketing Department, Wharton School, University of Pennsylvania, Philadelphia, USA. [14]Department of Psychology, University of Virginia, Charlottesville, USA. [15] Institute for Advanced Study, Massey University, Auckland, New Zealand.

[*] To whom correspondence should be addressed:

Anna Dreber
Department of Economics, Stockholm School of Economics
Box 6501, SE-113 83 Stockholm, Sweden.
Email: anna.dreber@hhs.se

Abstract

We implemented a decision market on 41 systematically selected *MTurk* social science experiments published in *PNAS* between 2015 and 2018. Social scientists (n = 162) traded on the outcome of direct replications, knowing that the 12 studies with the lowest and the 12 with the highest final market prices would be selected for replication, along with two randomly selected studies. The replication rate was 83% for the top-12 and 33% for the bottom-12 group. The correlation between decision market prices and replication outcomes was 0.505, indicating some potential for decision markets to select studies for replication. Overall, 54% of the studies were successfully replicated, with replication effect sizes averaging 45% of the original effect sizes. These results suggest that the replicability of *MTurk* experiments is comparable to that of previous systematic replication projects involving laboratory experiments. All study aspects were preregistered, and we strictly adhered to our pre-analysis plan.

Main

Can published research findings be trusted? Unfortunately, the answer to this question is all but straightforward, and the credibility of scientific findings and methods has been questioned repeatedly[1–9]. A vital tool for evaluating and enhancing the reliability of published findings is to carry out replications, which can be used to sort out true positive findings from false positives. In recent years, several systematic large-scale replication projects in the social sciences have been published[10–15]. While the results of these studies have generally been discouraging—with replication rates around 50% for findings previously reported as statistically significant in the literature—, the studies have substantially increased the interest in independent replications[16]. However, as it is time-consuming and costly to conduct replications, there is a need for a mechanism to decide which replications to prioritize to facilitate efficient and effective usage of resources[16–23]. In this study, we test the feasibility of one potential method to select which studies to replicate. Building on previous work using prediction markets to forecast replicability[14,15,24,25], we adapt this methodology into what is referred to as decision markets[26–29]. Decision markets are similar to prediction markets[30–32], but differ in that the market outcomes determine which studies to replicate.

To provide a "proof of concept" of using a decision market as a mechanism to determine which studies to replicate, we first identified all social science experiments published in the *Proceedings of the National Academy of Sciences* (*PNAS*) between 2015 and 2018 that fulfilled our inclusion criteria for *(i)* the journal and period; *(ii)* the platform on which the experiment was performed (*Amazon Mechanical Turk*; *MTurk*); *(iii)* the type of design (between-subjects or within-subject treatment design); *(iv)* the equipment and materials needed to implement the experiment (the experiment had to be logistically feasible for us to implement); and *(v)* the results reported in the experiment (that there is at least one statistically significant $p < 0.05$ main or interaction effect in the main text). Based on our inclusion criteria, we identified 44 articles, three of which have been excluded due to a lack of feasibility, leaving us with a final sample of 41 articles[28–68] (see Methods for details on the inclusion criteria). For each of these articles, we identified one critical finding with $p < 0.05$ that we could potentially replicate (see Methods for details and Supplementary Table 1 for the hypotheses tested in the 41 studies).

We then invited social science researchers to participate as forecasters in both a prediction survey and an incentivized decision market on the 41 studies. In the survey, the forecasters independently estimated the probability of replication for the 41 studies. In the decision market, they could trade on whether the result of each of the 41 studies would replicate. Participants in the decision market received an endowment of 100 tokens corresponding to USD 50, and 162 participants made a total of 4,412 trades. Traders in the market were informed about the preregistered decision mechanism: The 12 studies with the highest and the 12 studies with the lowest market prices were to be selected for replication; additionally, two randomly selected studies (out of the remaining 17 studies) are replicated to ensure incentive compatibility (see Methods for details).

Replication sample sizes were determined to have 90% power to detect ⅔ of the effect size reported in the original study at the 5% significance level in a two-sided test (with the effect size estimates having been converted to Cohen's *d* to have a common standardized effect size measure across the original studies and the replication studies; see Methods for details). If sample size calculations led to replication sample sizes smaller than in the original study, we targeted the same sample size as in the original study. The average sample size in the replications (n = 1,018) was 3.5 times as large as the average sample size in the original studies (n = 292). All replications were carried out online at *MTurk* as in the original studies.

Prior to starting the survey data collection (that preceded the decision market and replications), we preregistered[74,75] an analysis plan ("replication report") for each of the 41 potential replications at OSF after obtaining feedback from the original authors (https://osf.io/sejyp). After the replications had been conducted, the replication reports of the 26 studies selected for replication were updated with the results of the replications (and potential deviations from the protocol) and were posted to the same OSF repository. We also preregistered an overall pre-analysis plan at OSF before starting the data collection, detailing the study's design and all planned analyses and tests (https://osf.io/xsp6g). Unless explicitly stated, all analyses and tests reported in the manuscript have been preregistered and adhere exactly to our pre-registered analysis plan. Section 4 in the Supplementary Methods details any deviations from the planned design and analyses for the 26 replications. We preregistered two primary replication indicators and two primary hypotheses. The two primary replication indicators are the relative effect size of the replications and the statistical significance indicator for replication (i.e., whether or not the replication results in a statistically significant effect with $p < 0.05$ in the same direction as the original effect), which was the replication outcome predicted by forecasters in the survey and the decision market. In our two primary hypotheses, we conjecture that *(i)* the decision market prices positively correlate with the replication outcomes and *(ii)* the standardized effect sizes in the replications are lower than in the original studies. All hypotheses are evaluated using two-tailed tests, and—following Benjamin et al.[76]—we interpret results with $p < 0.005$ as "statistically significant evidence," whereas results with $0.005 \leq p < 0.05$ are considered "suggestive evidence."

Results

Replication outcomes and decision market performance. Fig. 1 and Supplementary Table 2 show the results for the decision markets where the final market price can be interpreted as the predicted replication probability. The predicted probabilities of replication range from 20.9% to 92.9% for the 41 studies, with a mean of 57.6% (sd = 23.6%). The average predicted probability for the 26 studies eventually selected for replication is 58.5%, suggesting that the studies selected for replication are "representative" of all the 41 studies in terms of the predicted replication rate. Fig. 1 also delineates the replication outcomes based on the statistical significance indicator, which allows for gauging the relationship between the decision market

prices and the replication outcomes. In Fig. 2 and Supplementary Table 3, we show the replication results for the 26 studies selected for replication. Of the 26 studies, 14 (53.8%; 95% CI [33.4%, 73.4%]) replicated successfully according to the statistical significance indicator. The point-biserial correlation between decision market prices and the binary replication outcome, testing our first primary hypothesis, is $r = 0.505$ (95% CI [0.146, 0.712]; $t(24) = 2.867$, $p = 0.008$; $n = 26$). Thus, in support of our first primary hypothesis, we find suggestive evidence of a positive association between decision market prices and replication outcomes. As a related secondary hypothesis, we test if the replication rate is lower among the 12 studies with the lowest decision market prices than for the 12 studies with the highest decision market prices. The replication rate is 33.3% (95% CI [9.9%, 65.1%]) for the studies in the "bottom-12" group and 83.3% (95% CI [51.6%, 97.9%]) for the studies in the "top-12" group, yielding suggestive evidence also in support of our secondary hypothesis test (Fisher's exact test; $\chi^2(1) = 6.171$, $p = 0.036$; $n = 24$).

Relative effect sizes. The mean effect size of the 26 replication studies (in Cohen's *d* units) is 0.253 ($sd = 0.357$) compared to 0.563 ($sd = 0.426$) for the original studies, implying a relative average effect size of 45.0%; the difference in effect sizes is statistically significant, supporting our second primary hypothesis of systematically smaller effect sizes in the replications (Wilcoxon signed-rank test; $z = 4.203$, $p < 0.001$; $n = 26$). The relative effect size can also be estimated for each study separately and varies between −17.0% and 136.2%, with a mean of 41.1% (95% CI [24.5%, 57.7%]). For the 14 studies that replicated according to the statistical significance indicator, the first and the second relative effect size measures as defined above are 69.5% and 72.0% (95% CI [54.8%, 89.3%]), indicative of some inflation in original effect sizes also for true positives. The two relative effect size measures for the 12 studies that failed to replicate according to the statistical significance indicator are 3.2% and 5.0% (95% CI [−2.6%, 12.5%]), respectively. Fig. 3 illustrates the relationship between the original and replication effect sizes.

Secondary replication indicators. We also preregistered four secondary replication indicators: the small-telescopes approach[77], the one-sided default Bayes factor[78], the replication Bayes factor[79], and the fixed-effects weighted meta-analytic effect size (see Methods for details). When relying on the small-telescopes approach, testing if the replication effect size is smaller than a "small effect,"[77] 15 studies (57.7%; 95% CI [36.9%, 76.6%]) are considered successful replications (Supplementary Figure 1 and Supplementary Table 4). The one-sided default Bayes factor (BF_{+0}) indicates the strength of evidence in favor of the alternative hypothesis as opposed to the null hypothesis. BF_{+0} exceeds one for the 14 studies (53.8%; 95% CI [33.4%, 73.4%]) that replicated according to the statistical significance indicator, with strong evidence ($BF_{+0} > 10$) for the tested hypothesis for nine studies (34.6%; 95% CI [17.2%, 55.7%]); BF_{+0} is below 1 for the 12 replications (46.2%; 95% CI [26.6%, 66.6%]) that failed to replicate according to the statistical significance indicator, with strong evidence ($BF_{+0} < 0.1$) for the null hypothesis for seven studies (26.9%; 95% CI [11.6%, 47.8%]) based on the evidence categories proposed by

Jeffreys[80] (Supplementary Figure 2 and Supplementary Table 4). The one-sided replication Bayes factor (BF_{R0}) indicates the strength of additional evidence in favor of the alternative hypothesis as opposed to the null hypothesis, given the already acquired evidence based on the original data set[79]. Replication Bayes factors lead to similar conclusions as the one-sided default Bayes factors, with $BF_{R0} > 10$ for ten studies (38.5%; 95% CI [20.2%, 59.4%]) and $BF_{R0} < 0.1$ for seven studies (26.9%; 95% CI [11.6%, 47.8%]) one exception to this is the study by Cooney et al.[43], for which the default Bayes factor exceeds one ($BF_{+0} = 8.01$) but the replication Bayes factor is below one ($BF_{R0} = 0.23$) due to the replication effect size being only about a third of the original effect size and a larger sample size in the replication compared to the original study (Supplementary Figure 2 and Supplementary Table 4). The meta-analytic effect size is statistically significant at the 5% level for 16 studies (61.5%; 95% CI [40.6%, 79.8%]) and significant at the 0.5% level for 14 studies [53.8%; 95% CI [33.4%, 73.4%]); see Supplementary Figure 3 and Supplementary Table 4. The meta-analytic effect sizes should be interpreted cautiously as the original effect sizes are likely to be overestimated on average due to insufficient statistical power (and potentially due to questionable research practices)[81,82]. Overall, the various replication indicators yield congruent binary conclusions for 23 of the 26 replications.

Survey forecasts vs. decision market predictions. We tested three additional preregistered secondary hypotheses based on the survey beliefs about replication (see Supplementary Table 5 for the survey results). The point-biserial correlation between average survey beliefs and the replication outcomes based on the statistical significance criterion is $r = 0.476$ (95% CI [0.107, 0.694]; $t(24) = 2.650$, $p = 0.014$; $n = 26$). The survey beliefs and the decision market prices are—as expected—positively correlated with a Pearson correlation of 0.899 (95% CI [0.814, 0.944]; $t(39) = 12.830$, $p < 0.001$; $n = 41$) (Fig. 4a). The final secondary hypothesis tests if the prediction accuracy, measured in terms of the absolute prediction error and the Brier score (i.e., the squared prediction error), is higher for the decision market than the survey (Fig. 4b). The mean absolute prediction error and the mean Brier score are 0.353 and 0.188 for the decision market, and 0.421 and 0.202 for the survey, respectively, implying suggestive evidence for higher accuracy for the market forecasts based on the absolute prediction error (Wilcoxon signed-rank test: $z = 2.172$, $p = 0.030$; $n = 26$) but not the Brier score (Wilcoxon signed-rank test: $z = 1.181$, $p = 0.238$; $n = 26$).

Beliefs about the impact of the Covid-19 pandemic on replicability. A potential issue raised by some original authors in giving feedback on the replication reports prior to the data collection was that the replicability of some original results might be affected by the implications of the Covid-19 pandemic (as the original studies were conducted before the pandemic). We evaluate this possibility in a preregistered exploratory analysis, relying on the forecasters' beliefs about the impact of the pandemic on replicability. As part of the prediction survey, participants were asked to judge whether the pandemic would have affected the likelihood of successful replication, measured on a scale from −3 ("the pandemic has definitely decreased the probability of replication") to 3 ("the pandemic has definitely increased the probability of

replication"). We test if the average response to this question differs from zero using a one-sample *t*-test for each of the 26 replications, and we test if the average response across all 26 studies differs from 0. We find a statistically significant result for four and a suggestive result for two replications on beliefs that Covid-19 has affected the replication probability (Supplementary Table 5). For the six studies with suggestive or statistically significant evidence, the estimate is negative for two studies and positive for four; only in two of the cases does the sign of the expectation match the eventual replication outcome. For the average belief about the impact of the pandemic on replicability across the 26 studies of 162 forecasters (who were active in the decision markets), there is suggestive evidence that the mean of 0.039 (*sd* = 0.190) differs from zero ($t(161) = 2.598$, $p = 0.010$; $n = 162$). Somewhat surprisingly—and in contrast to the concerns raised by some of the original authors—there is thus a tendency for forecasters to believe that the pandemic has *increased* the average likelihood that the studies will replicate. However, the magnitude of the effect is small ($d = 0.204$; 95% CI [0.049, 0.360]). In addition, we tested, estimating the point-biserial correlation, if the average belief (per study) about the pandemic's impact on replicability correlates with the replication outcomes based on the statistical significance indicator; we do not find a statistically significant association ($r = 0.014$, 95% CI [−0.360, 0.382]; $t(24) = 0.068$, $p = 0.946$; $n = 26$). The non-significant and close-to-zero correlation provides some reassurance that experts from the field, on average, do not expect that the pandemic would have substantially compromised the replication results. Yet, we cannot rule out that Covid-19 has entailed effects on replicability not foreseen by the scholars participating in the survey. Forecasters' beliefs about the pandemic's impact on replicability are also neither statistically significantly correlated with the final decision market prices ($r = 0.387$, 95% CI [−0.008, 0.669]; $t(24) = 2.055$, $p = 0.051$; $n = 26$) nor the average survey belief of replication ($r = 0.347$, 95% CI [−0.053, 0.644]; $t(24) = 1.815$, $p = 0.082$; $n = 26$), although the point estimates of the correlations are quite sizeable.

Discussion

Replications are essential to assess and enhance the credibility of scientific claims, but replications are costly. A mechanism to determine which claims to prioritize is needed to allocate replication resources efficiently. This study serves as a proof of concept for one potential mechanism: decision markets.

We found suggestive evidence ($p < 0.05$) for our first primary hypothesis that final decision market prices correlate with replication outcomes, $r = 0.505$. This result suggests that decision markets show some potential as a mechanism to determine which findings to prioritize for replication. However, the observed effect size is somewhat smaller than the effect size of $r = 0.67$ presumed in our a priori power calculations (see Methods for details). The correlation is within the range of previous prediction markets on systematic replication projects with correlations of 0.42 in the *Replication Project: Psychology* (*RPP*)[13,24], 0.30 in the *Experimental Economics Replication Project* (*EERP*)[14], and 0.84 in the *Social Sciences Replication Project*

(*SSRP*)[15], but we expected a stronger correlation because we selected studies with the highest and the lowest prices for replication. Consistent with the primary hypothesis test, there is also suggestive evidence of a difference in the replication rate between the “top-12” (10 of 12) and “bottom-12” (4 of 12) in our secondary hypothesis test. The difference of 50 percentage points is also reflected in the difference between the forecasted replication rates of 86.6% ("top-12") vs. 29.6% ("bottom-12") in the decision market. However, the small sample size suggests caution against drawing firm conclusions about whether decision markets are appropriate for selecting studies for replication.

The pooled evidence from previous prediction market studies on replication outcomes suggests that markets are somewhat more accurate than surveys[25], although the difference tends to be small. These indications are consistent with our results, yielding suggestive evidence of higher accuracy in terms of the absolute prediction error but not in terms of the squared prediction error (Brier score). The correlation between the average survey beliefs and the replication outcomes was almost as high for the survey as the prediction market (0.476 vs. 0.505). Since surveys are less resource-intensive, simple polls can be an expedient alternative to decision markets for selecting which studies to replicate, even if they should be somewhat less accurate. However, a caveat in interpreting the prediction accuracy of surveys in this context is that participants may put more effort into the survey, knowing that they will later participate in an incentivized decision or prediction market. Another potential method for selecting which studies to replicate would be to rely on the original *p*-value for studies reporting statistically significant results[25]. The point-biserial correlation (not pre-registered) between the original *p*-value and the replication outcome is −0.400 ($p = 0.043$; 95% CI [0.014, 0.648]) and comparable in magnitude to correlations of −0.33 in the *RPP*[13], −0.57 in the *EERP*[14], and −0.40 in the *SSRP*[15]. Although the prediction accuracy appears to be somewhat lower for original *p*-values than market and survey forecasts[25], relying on *p*-values may well be considered a practical alternative as it does not involve any additional data collection. Another possibility would be to use predicted replication probabilities from machine learning models to select studies for replication. There has been some progress in developing such models[83–86], but evidence on whether they outperform markets or surveys is yet missing. Other potential mechanisms for selecting which studies to replicate include relying on general or study-specific characteristics (e.g., connection to theory, surprise factor, sample size, effect size)[16], relying on cost-benefit considerations[17,18], employing Bayesian strategies[19], determining the “replication value”[20], adopting empirical audit and review[21], or using predictions from laypeople[22,23].

Using decision markets to select the studies with the highest and lowest predicted probabilities for replication is just one of many potential selection rules for this methodology. Our goal was to test whether a decision market could separate true positives from false positives, and we aimed to maximize the statistical power of detecting an association between market prices and replication outcomes. For the practical application of decision markets, the choice of the selection mechanism will largely depend on the objective function. One selection rule would be

to choose the studies with the highest predicted false positive likelihood, i.e., the studies with the smallest market prices (in addition to randomly selected studies to ensure incentive compatibility). This decision mechanism would align with the objective of identifying and correcting false discoveries in the literature to facilitate an efficient allocation of resources for follow-up investigations. Another selection rule would be to replicate the studies with market predictions close to 50%, which reflects the highest possible uncertainty or disagreement regarding the likelihood of the original finding being genuinely true. Providing additional evidence on these claims could maximize the information value of replication studies, as well-powered replications will move the probability that the tested hypothesis is genuinely true towards 0% or 100%.

For our second primary hypothesis, we found strong evidence that original effect sizes are inflated on average compared to replication effect sizes, with a relative average effect size of 45%. This is comparable to previous systematic replication studies, with relative average effect sizes of 49% in the *RPP*[13], 59% in the *EERP*[14], and 54% in the *SSRP*[15]. The replication rate of 54% based on the statistical significance indicator is also similar to previous replication studies, with 36% in RPP[13], 61% in the EERP[14], and 62% in the SSRP[15]. The ability of the statistical significance indicator to discriminate between true positives and false positives depends on replication power, and the relative average effect size of the studies that failed to replicate should be close to zero if the systematic replication study successfully separates false positives from true positives. The relative average effect size of the 12 studies that failed to replicate according to the statistical significance indicator was 3.2%, which is close to zero and consistent with a successful separation between true positives and false positives. But also true positive findings can be expected to have exaggerated effect sizes in the published literature due to a lack of statistical power[81,82]. In line with this, we found a relative average replication effect size of 69.5% for the 14 studies that successfully replicated based on the statistical significance indicator. These findings are consistent with similar analyses in the *SSRP*[15] in which the mean relative effect size among the studies that failed to replicate according to the statistical significance indicator was 0.3%, and the mean relative effect size among the studies that replicated successfully was 73.1%. This illustrates how the combination of statistical significance and relative effect size can contribute to revealing possible false positives and true positives with exaggerated effect sizes.

Previous systematic replication studies have focused on laboratory experiments rather than online experiments. Concerns have been raised over data quality in online data collections using “crowd workers,” as via *MTurk*[87–93], and part of the rationale for zeroing in on experiments conducted via *MTurk* was that we tend to share these concerns. However, the results of this study do not suggest that replicability is substantively lower for experiments conducted via *MTurk* compared to experiments conducted in physical laboratories for studies published in top journals; more evidence is needed to draw strong conclusions. Relatedly, the predicted average replicability rate of 57.6% in the decision market—despite widespread concerns about data

quality on *MTurk*—is within the range of replication rate forecasts in previous prediction markets of 56% in the *RPP*[24], 75% in the *EERP*[14], and 63% in the *SSRP*[15]. We used IP quality checks[89,94] to minimize the chances of low-quality participant data (see Methods for details), screening out participants before the random assignment into treatments. In total, across all 26 replications, 29% of the participants who accepted a "human intelligence task" (HIT) failed the IP check and were excluded (this descriptive result was not preregistered; see Methods for further details). The replication results from our study should thus not be extrapolated to *MTurk* experiments not using a comparable screening procedure.

A successful replication, on its own, does not provide valid evidence for the tested hypothesis. A finding can be replicable and be based on an invalid experimental design, leading to biased results. An example of this would be an online design that systematically results in more attrition in one experimental treatment, causing selection bias in favor of the tested conjecture.[87] Likewise, a failed replication, on its own, does not provide direct evidence against the tested hypothesis. A finding can be unreplicable and based on an invalid experimental design, leaving the hypothesis untested. Although the replication rate for online experiments in our study appears to be similar to previous laboratory evidence, it does not necessarily imply that online and laboratory experiments provide equally valid evidence of the tested hypotheses.

In this proof-of-concept investigation of decision markets for assessing replicability, decision markets show potential as a tool for selecting studies for replications, but further work is needed to draw strong conclusions. We also observe that the replication rate of social science experiments based on data collections via *MTurk* published in *PNAS* is comparable to previous systematic replication projects of experimental studies in the social sciences, primarily based on lab experiments.

Methods

We preregistered an analysis plan for the project at OSF on October 7, 2021, prior to starting the survey data collection (that preceded the decision market and replications), which detailed the design of the study and the exact analyses for all planned analyses and tests (https://osf.io/xsp6g). Unless explicitly mentioned in the main text, we adhere exactly to our pre-analysis plan. The information in this Methods section follows the pre-analysis plan (PAP; with some of the information from the pre-analysis plan reported in the Supplementary Methods, Sections 1–3).

Prior to starting the survey data collection, we also preregistered an analysis plan ("replication report") for each of the 41 potential replications included in the decision market at OSF after obtaining feedback from the original authors (https://osf.io/sejyp). After the replications had been conducted, the 26 replication reports of the replications selected for replication by the decision market were updated with the results of the replications and posted at the same OSF repository. Any deviations from the preregistered analysis plans for the 26 replications are detailed in the

26 “post-replication reports” and listed in Supplementary Methods, Section 4. We provided all original authors the opportunity to comment on the replication results (without a particular due date) and make the comments publicly available as we receive them alongside the post-replication reports on OSF (https://osf.io/sejyp).

Below, we provide further details on the inclusion criteria, the decision market setup and the survey, the replications, and the replication rate indicators included in the study. The preregistered analyses and tests were divided into descriptive results of the replication rate among the 26 replicated studies and hypothesis tests. The preregistered descriptive results were furthermore divided into primary replication indicators and secondary replication indicators, and the pre-registered hypothesis tests were divided into *(i)* primary hypotheses, *(ii)* secondary hypotheses, and *(iii)* exploratory analyses. See Supplementary Methods, Section 3, for more details about the preregistered hypothesis tests and exploratory analyses.

Inclusion criteria for studies

We reviewed all *PNAS* articles from 2015–2018 and searched for the terms *Amazon Mechanical Turk*, *MTurk*, and *Turk*. We included all social sciences articles that fulfilled our inclusion criteria for *(i)* the journal and time period, *(ii)* the platform on which the experiment was performed (*MTurk*), *(iii)* the type of design (between-subjects or within-subject treatment design), *(iv)* the equipment and materials needed to implement the experiment (the experiment had to be logistically feasible for us to implement), and *(v)* the results reported in the experiment (that there is at least one statistically significant $p < 0.05$ main or interaction effect in the main text). Based on the inclusion criteria, we identified 44 articles. After contacting the original authors, we ended up with 41 articles (the three excluded articles[95–97] involved either software or platforms that no longer existed or methods we were unfamiliar with). In these 41 articles, we identified at least one critical finding that we could replicate. In cases where several studies in the same article fit the inclusion criteria, we randomly picked one of the studies; this was the case for 17 of the 26 replicated studies (Ames and Fiske[33], Atir and Ferguson[34], Baldwin and Lammers[35], Boswell et al.[37], Cooney et al.[43], Genschow et al.[46], Gheorghiu et al.[47], Halevy and Halali[49], Hofstetter et al.[52], John et al.[56], Jordan et al.[57], Klein and O’Brien[60], Kouchaki and Gino[61], McCall et al.[63], Rai et al.[68], Stern et al.[71], and Williams et al.[73]). In cases where the (randomly picked) study contained several conditions, we randomly picked which to compare to the control condition. After that, we looked for the central result with $p < 0.05$ for that particular study. If there were several statistically significant results, one was selected at random. Needless to say, the replication results only pertain to the single central result selected per paper, and the replication outcome does not necessarily generalize to other results reported in the original articles. For convenience, we refer to the replications as “replication of [study reference],” though.

For Cheon and Hong[41], the result chosen for replication is reported as part of a 2x2 ANOVA in the original article; since the paper does not report the main effect, the original authors kindly provided us with the corresponding estimates. For Gheorghiu et al.[47], the result to be replicated is only reported with its *p*-value in the paper; a precise estimate of the test statistic has been obtained from a re-analysis of the original data, which the original authors kindly provided. For the study by Kraus et al.[62], we could not reproduce the result reported in the original article using the original data. The original authors acknowledged that there had been a reporting error in the original article. For the replication, we use the analysis described in the paper; the effect size and the test statistic reported in the original paper were replaced by the re-estimated result. For the study by Williams et al.[73], the focal hypothesis test in the replication is based on a composite score of five suites of behavior (which are tested separately in the original article) to have a single test. The original authors report tests on composite measures in the Supporting Information of their article too, and they approved the choice to investigate the replicability of the focal hypothesis using a composite score. These changes are transparently reported in the replication reports for each study (see https://osf.io/sejyp for details).

Decision market and prediction survey

We invited researchers to voluntarily participate in the decision market through public mailing lists (ESA and JDM lists) and social media (e.g., Twitter/X); we also emailed colleagues asking them to distribute the call for participants within their professional networks. Participants were required to hold a Ph.D. degree or to be a Ph.D. student currently. In the decision market, participants bet on whether or not the specific result chosen for each study would replicate based on the statistical significance indicator ($p < 0.05$ in the replication and an effect in the same direction as in the original study) as a criterion for replication (thus a binary outcome, as discussed below). Prior to the decision market, participants filled out a survey where we asked them to assign a probability of successful replication to each of the 41 results. The survey is available at https://osf.io/a24zq. Completing the survey was a prerequisite for participating in the markets. We started the recruitment of participants for the decision market on October 4 (2021), and we started sending out the prediction survey on October 8 to those who had signed up for the study (participants who signed up after October 8 received the survey invitation a few days after their registration). The deadline for registering as a participant was October 29, and the deadline for completing the survey was November 5. Overall, 289 participants signed up to participate and were forwarded the link to the survey; 193 participants started the survey, and 162 completed it by the due date.

In the survey, we asked participants to assess, for each replication study, (i) the likelihood that the hypothesis will successfully replicate (on a scale from 0% to 100%); (ii) their stated expertise for the study/the hypothesis (on a scale from 1 to 7); and (iii) whether they believe the pandemic has affected the likelihood of replication. The question about the pandemic was measured on a scale from −3 (“the pandemic has definitely decreased the probability of replication”) to 3 (“the

pandemic has definitely increased the probability of replication"); the 0 midpoint implies that they do not think that the pandemic has affected the probability of replication. The survey concluded with some (optional) demographic questions. The survey was not incentivized.

The decision market opened on November 8 (2021) and closed after two weeks on November 22 (and before the decision market opened, participants had at least one week to complete the prediction survey). In the decision market, participants could trade (bet) on whether they expected the 41 studies to replicate. While participants had the opportunity to bet on the replication outcome of the 41 studies, we did not carry out replications for all 41 studies, but the final decision market prices determined which studies to replicate. We replicated the 12 studies that had the highest and the 12 studies that had the lowest market prices when the market closed. Additionally, two out of the remaining 17 studies were randomly selected for replication to ensure a non-zero probability for each study to be replicated (i.e., we replicated 12 + 12 + 2 = 26 studies in total). The decision rule for which studies to replicate was based on final market prices and was common knowledge to the market participants; the instructions (provided to participants who completed the prediction survey by the due date) are available at https://osf.io/a24zq/.

We chose 12 studies with the lowest predicted probability and 12 studies with the highest predicted probability based on a power calculation using the pooled data from our previous prediction market studies[25]. The power calculations were conducted by randomly sampling 41 studies from the dataset described in Gordon et al.[25] in a simulation with 10,000 iterations and then selecting the forecasts and outcomes from the 12 studies with the lowest predicted probability, the 12 studies with the highest predicted probability, and two random studies. We failed to set a random seed for the simulation when the study was conducted, implying that the pre-registered power estimates could not be numerically reproduced when we wrote up the study results. For full transparency, we report the power estimates included in the PAP in parentheses below for full transparency. The median point-biserial correlation coefficient across the 10,000 runs is 0.671 (reported as 0.66 in the PAP), and we have 91.0% power (reported as >90% in the PAP) to detect a statistically significant correlation (n = 26) between the decision market prices and the replication outcomes at the 0.5% level and 99.4% power (reported as >95% in the PAP) to detect a statistically significant correlation at the 5% level, which is our first primary hypothesis test. As a secondary hypothesis test, we test if the fraction of studies that successfully replicate differs between the 12 studies with the highest and the 12 studies with the lowest predicted replication probabilities using Fisher's exact test. Applying the same sampling approach as for primary hypothesis 1, the median difference in replication rates between the 12 studies with the highest and the 12 studies with the lowest market prices is 0.663; the secondary test (n = 24) has 66.5% power at the 0.5% level (reported as 66% in the PAP) and 94.9% power at the 5% level (reported as 95% in the PAP). The code for the power simulations is available at https://osf.io/47drs.

Implementation of the decision market. We used a web-based trading platform, similar to the ones used in Camerer et al.[14,15] and identical to the one used in Botvinik-Nezer et al.[6]. The trading platform involves two main views: (i) the market overview and (ii) the trading page. The market overview listed the 41 assets (i.e., one corresponding to each study) in tabular format, including information on the current price for buying a share and the number of shares held (separated for long and short positions). Via the trading page, which was shown after clicking on any of the assets, participants could make investment decisions (i.e., buy or sell shares) and view price developments in graphical format for the particular asset.

Trading and incentivization. Decision market participants received an endowment of 100 tokens corresponding to USD 50. Once the markets opened, market participants could use the tokens to trade shares of the assets available in the market. An automated market maker, implementing a logarithmic market scoring rule[98], determined the assets' prices. At the beginning of the markets, all assets were valued at 0.50 tokens per share. The market maker calculated the share price for each infinitesimal transaction and updated the price based on the scoring rule. With this mechanism, participants had incentives to invest according to their beliefs[28,29]. With the logarithmic scoring rule, the price p for an infinitesimal trade is determined as $p = e^{s/b} \div (e^{s/b} + 1)$, where s denotes the net sales (shares held − shares borrowed) that the market maker has done so far in a market; the liquidity parameter b determines how strongly the market price is affected by trade and was set to $b = 100$, implying that by investing ten tokens, traders could move the price of a single asset from 0.50 to about 0.55. Decision market participants were paid only for studies chosen for replication (based on their final holdings). Participants received one token per correct share for the replications with the 12 lowest and 12 highest final market prices. For the two randomly selected replications, participants received $17 \div 2 = 8.5$ tokens for each share; for replications that were not selected for replication, participants received no compensation for their holdings. We followed this procedure to keep information revelation in the decision market incentive-compatible, with the increased payouts for the randomly selected studies compensating for the "voided" shares in studies not selected for replication. Participants were paid after all 26 replications had been completed.

Participation. A total of 193 participants completed the prediction survey (a prerequisite to participate in market trading) and were subsequently invited to trade on the decision market. Of these 193 participants, 162 (83.9%) traded in the market at least once. During the two-week trading period, a total of 4,412 transactions were recorded. On average, each trader prompted 27.2 transactions ($sd = 30.7$; min = 1, max = 185). The average number of traders per hypothesis was 65.1 ($sd = 15.3$; min = 35, max = 98); the average number of transactions recorded per hypothesis was 107.6 ($sd = 35.2$; min = 56, max = 213). See Supplementary Table 2 for descriptive statistics on the trading activity for each market.

Replications

The replications started in January 2022 and were completed in October 2023. The replications were planned and pre-registered by five replication teams: a team at CalTech, LMU, and Wharton; a team at the Stockholm School of Economics; a team at the National University of Singapore; a team at the University of Amsterdam; and a team at the University of Innsbruck.

Participants in replication studies. All replications were carried out at Amazon Mechanical Turk as in the original studies. We ensured that participants could only participate once using the same account in a specific study. If the original study had not specified a HIT approval rate, we recruited participants with a HIT approval rate of at least 95%; if the original study had specified a higher approval rate, we applied the same requirement as used in the original study.

Before forwarding participants to each study, we forwarded the IP addresses to https://www.ipqualityscore.com/ for a quality check to minimize the chances of low-quality participant data (we initially planned to use this filter ex-post but during the data collection of the first two replication studies of Klein & O'Brien[60] and Halevy & Halali[49] we decided to set it up so that the IP address quality check happened before participants got redirected to the study). Participants for whom one or more of the following was true could not proceed with participating in the study: fraud score ≥ 85; TOR = True; VPN = True; Bot = True; abuse velocity = high. This means that, for example, participants were not allowed to use a virtual private network (VPN) or Tor connections or participate if they had IP addresses that had recently engaged in automated bot activity (the VPN exclusions were made ex-ante, i.e., before participants were redirected to the study, for 4 studies and ex-post for 22 studies). Thereafter, in all replications, participants were first shown a Captcha and then provided informed consent. After this, we included an attention check that participants had to pass to proceed to the study (with the exception of Reeck et al.[69]; see Section 4 in the Supplementary Information for details). The attention check was implemented in addition to any other potential attention check(s) used in the original study. All these exclusions based on the "quality filters" were preregistered, but the pre-analyis plan did not specify if participants would be excluded before or after participating in the study.

The individual replication studies sometimes also used additional exclusion criteria that are detailed in the preregistered replication report for each replication (we tried to use the same exclusion criteria for the replications as used in the original studies as much as possible). The replication sample sizes defined below are the sample sizes after any exclusions of participants.

Replication sample sizes. The replications were carried out with high statistical power. Replication sample sizes were based on having 90% power to detect ⅔ of the effect size reported in the original study (with the effect size converted to Cohen's *d* to have a common standardized effect size measure across the original studies and the replication studies). See Supplementary Methods, Section 1, for more details about the power calculations and replication sample sizes. The criteria for replication were an effect in the same direction as the original study and a p-value < 0.05 (in a two-sided test). In cases where this power estimation

led to a sample size smaller than the original one, we used the same sample size as in the original study. The average replication sample ($\bar{n}$ = 1,018) size was 3.5 times as large as the average sample size of the original studies ($\bar{n}$ = 292). We continued the data collection for each replication until we reached at least the preregistered sample size after exclusions for that replication, and this led to slightly larger replication sample sizes than preregistered in all replications except one (as it is not possible with exclusion criteria to get an exact sample size as the number of exclusions is not known ex-ante).

Conversion of effect sizes to Cohen's d. We converted the effect sizes of all the original studies and all the replication studies to Cohen's *d* to have a standardized effect size (the effect size in the original study was always assigned a positive sign; the effect size in the replication study was assigned a positive sign if the effect was in the same direction as in the original study and a negative sign if the effect was in the opposite direction of the original study). See Supplementary Methods, Section 2 for details about the conversion of effect sizes to Cohen's *d*.

Replication reports. For each of the 41 studies, we prepared a pre-replication plan/report stating the hypothesis we had chosen from each paper and how we planned to proceed with the replication study. These reports were shared with the original authors for feedback, and at least one original author from each paper replied. These pre-replication reports were posted at OSF (https://osf.io/sejyp at the same time as the pre-analysis plan and prior to the start of the prediction survey (that preceded the decision markets and the replication data collections). For those studies that were selected for replication, we have updated the replication reports with the replication results after the replications were completed. After sharing them with the original authors for feedback, we have posted the updated replication reports at OSF as well (https://osf.io/sejyp). Additionally, we reached out to the original authors for their comments on the replication reports and results. We promised to make their comments available along with the replication reports, and any comments received can be found at https://osf.io/sejyp.

Subject payments. We standardized payments across all replications such that studies had a certain show-up fee depending on the expected length of the study. In particular, we paid an hourly fee of USD 8.00 for all studies, and we calculated the show-up fee for each study based on the expected length of the study. For all studies, we implemented a minimum payoff of USD 1.00. For studies with incentive payments, we used the same incentive payment as in the original study, paid in addition to the show-up fee. If we faced problems in recruiting participants, we increased the show-up fee, which happened for two studies[48,52].

Replication indicators

Statistical significance criterion (primary replication indicator). The first primary replication indicator was the statistical significance criterion – i.e., whether the replication resulted in an effect size in the same direction as the original study and a two-sided *p*-value less than 0.05. Unless otherwise stated above, we used the same statistical test as in the original study. We

report the replication rate (i.e., the fraction of the 26 studies that replicated according to this criterion) and the 95% Clopper-Pearson CI of this fraction in the Results section. We also report the 95% CI of the replication effect size for each of the 26 replication studies in Fig. 2 (normalized such that the original effect size equals one) and Supplementary Table 3.

Relative effect sizes (primary replication rate indicator). As a second primary replication indicator, we used relative effect sizes. Relative effect sizes were estimated in two different ways. We report the mean effect size of all 26 replications and compare it to the mean effect size of the 26 original studies (see also primary hypothesis test 2 below). We furthermore estimate the relative effect size of each replication (the replication effect size divided by the original effect size) and estimate the mean of this variable for the 26 replication studies and the 95% CI of this mean (based on a one-sample *t*-test). We report both of these measures of the relative effect size separately for the replications that replicate and those that do not. These results are reported in the Results section, Fig. 3, and Supplementary Table 3.

Small-telescopes approach (secondary replication indicator). We also used the small-telescopes approach[77]. For this indicator, we estimated whether the replication effect size was significantly smaller (using a one-sided test at the 5% level) than a "small effect," defined as the effect size the original study would have had 33% power to detect. For studies using *t*-tests (or *F*-tests converted to a *t*-test statistic), we based "the small effect size" on the effect size that a *t*-test had 33% power to detect (at the 5% level in a two-sided test); and for studies using *z*-test statistics (or chi-square tests converted to a *z*-test statistic), we based "the small effect size" on the effect size that a *z*-test had 33% power to detect (at the 5% level in a two-sided test). To test whether the replication effect size was significantly smaller than "the small effect size" in a one-sided test at the 5% level, we estimated a 90% CI of the replication effect size. We tested if the 90% CI overlapped the small effect size, with CIs constructed as described in Supplementary Methods, section 2. If the effect size in the replication was significantly smaller than this "small effect size," the result was considered a failed replication; otherwise, it was considered successful. We report the fraction of studies that replicate according to this criterion and the 95% Clopper-Pearson CI of this fraction. The small-telescopes results are reported in Supplementary Figure 1 and Supplementary Table 4.

Bayes factors (secondary replication indicators). We also compute the one-sided default Bayes factors on the replication data, allowing us to 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 in terms of a truncated Cauchy distribution with scale 0.707 was assigned to the size of the effect) versus the null hypothesis that stipulates the effect to be absent[78]. In addition, we also computed (one-sided) replication Bayes factors, which quantifies the additional evidence for the hypothesis given the evidence already provided by the original study[79]. These results are reported in Supplementary Figure 2 and Supplementary Table 4. We use the evidence categories proposed by Jeffreys[80] to interpret the Bayes Factors. A detailed report on the estimation of the Bayes factors is available at https://osf.io/47drs/.

Meta-analytic effect sizes (secondary replication indicator). We estimated the meta-analytic estimate of the effect size by combining the original result and the replication result in a fixed-effect meta-analysis. We report the fraction of the 26 studies that replicated according to the 0.05 and the 0.005 significance threshold and the 95% Clopper-Pearson CI of these fractions. We also use the stricter 0.005 significance threshold as a replication indicator for the meta-analytic effect sizes because this is similar to observing two studies (an original study and a replication study) that are significant at the 0.05 level. We report these results in the Results section, Supplementary Figure 3, and Supplementary Table 4.

Ethical Approval

We sought ethical approval from the Swedish Ethical Review Authority who had no ethical objections to the decision market part of the project and judged the replication part of the project to not be covered by the Swedish ethical review law (Dnr 2019-06501).

Data and Code Availability

The data reported in this paper is tabulated in Supplementary Tables 1–5. The replication reports (both the pre-replication and the post-replication versions), the pre-analysis plan, the data from the survey and the decision market, the data for each of the 26 replications, and the analysis scripts generating all results, figures, and tables reported in the main text and the Supplementary Information are available at the project's OSF repository (https://osf.io/sk82q).

Acknowledgments

We thank Alexander Andevall for help with the data collection and programming of experiments and Robb Willer for helpful advice on defining the IP address check and exclusion criteria used to exclude individuals from participating to minimize low-quality participant data. For financial support, we thank the Austrian Science FWF (grant SFB F63), Jan Wallander and Tom Hedelius Foundation (grants P21-0091 and P23-0098), Knut and Alice Wallenberg Foundation and Marianne and Marcus Wallenberg Foundation (Wallenberg Scholar grant to A.D.), and Riksbankens Jubileumsfond (grant P21-0168).

Author Contributions

A.D., F.H., J.H., M.J., M.K., B.A.N., and T.P. designed the study; A.D., F.H., and M.J. managed the study; Y.C., A.D., F.H., M.J., and T.P. designed and implemented the decision market; A.D., F.H., M.J., B.M., and V.W. selected articles and critical findings for (potential) replication; A.D., C.F.C., F.H., T.-H.H., S.H., J.H., N.I., T.I., L.J., M.J., M.K., B.M., D.M., G.N., A.S., R.S., E.-J.W., and V.W., designed the replications and collected replication data; F.H, S.H., T.I., L.J., A.S., R.S., and V.W. conducted the preregistered statistical tests on the individual replications; A.L. computed the Bayes factors; F.H. conducted all analyses reported in the manuscript; A.D., F.H., and M.J. wrote the paper; all authors reviewed and approved the final manuscript.

Competing Interest Statement

The authors declare no competing interests.

References

1. Leamer, E. E. Let's take the con out of econometrics. *Am. Econ. Rev.* **73**, 31–43 (1983).
2. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
3. Begley, C. G. & Ellis, L. M. Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
4. McNutt, M. Reproducibility. *Science* **343**, 229–229 (2014).
5. Gertler, P., Galiani, S. & Romero, M. How to make replication the norm. *Nature* **554**, 417–419 (2018).
6. Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).
7. Breznau, N. *et al.* Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proc. Natl. Acad. Sci.* **119**, e2203150119 (2022).
8. Delios, A. *et al.* Examining the generalizability of research findings from archival data. *Proc. Natl. Acad. Sci.* **119**, e2120377119 (2022).
9. Huber, C. *et al.* Competition and moral behavior: A meta-analysis of forty-five crowd-sourced experimental designs. *Proc. Natl. Acad. Sci.* **120**, e2215572120 (2023).
10. Klein, R. A. *et al.* Investigating variation in replicability: A "many labs" replication project. *Soc. Psychol.* **45**, 142–152 (2014).
11. 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).
12. Klein, R. A. *et al.* Many Labs 2: Investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).
13. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
14. Camerer, C. F. *et al.* Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
15. Camerer, C. F. *et al.* Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).
16. Pittelkow, M.-M. *et al.* The process of replication target selection in psychology: What to consider? *R. Soc. Open Sci.* **10**, 210586 (2023).
17. Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M. & Lakens, D. The costs and benefits of replication studies. *Behav. Brain Sci.* **41**, e124 (2018).
18. Alipourfard, N. *et al.* Systematizing Confidence in Open Research and Evidence (SCORE). Preprint at https://doi.org/10.31235/osf.io/46mnb (2021).
19. Hardwicke, T. E., Tessler, M. H., Peloquin, B. N. & Frank, M. C. A Bayesian decision-making

framework for replication. *Behav. Brain Sci.* **41**, e132 (2018).

20. Isager, P. M. *et al.* Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychol. Methods* **28**, 438–451 (2023).
21. O'Donnell, M. *et al.* Empirical audit and review and an assessment of evidentiary value in research on the psychological consequences of scarcity. *Proc. Natl. Acad. Sci.* **118**, e2103313118 (2021).
22. Hoogeveen, S., Sarafoglou, A. & Wagenmakers, E.-J. Laypeople can predict which social-science studies will be replicated successfully. *Adv. Methods Pract. Psychol. Sci.* **3**, 267–285 (2020).
23. Marcoci, A. *et al.* Predicting the replicability of social and behavioural science claims from the COVID-19 Preprint Replication Project with structured expert and novice groups. Preprint at https://doi.org/10.31222/osf.io/xdsjf (2023).
24. Dreber, A. *et al.* Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl. Acad. Sci.* **112**, 15343–15347 (2015).
25. Gordon, M., Viganola, D., Dreber, A., Johannesson, M. & Pfeiffer, T. Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *PLoS One* **16**, e0248780 (2021).
26. Hanson, R. Decision markets. *IEEE Intell. Syst.* **14**, 16–19 (1999).
27. Hanson, R. Combinatorial information market design. *Inf. Syst. Front.* **5**, 107–119 (2003).
28. Chen, Y., Kash, I., Ruberry, M. & Shnayder, V. Decision markets with good incentives. in *Internet and Network Economics* (eds. Chen, N., Elkind, E. & Koutsoupias, E.) 72–83 (Springer, 2011). doi:10/b4vtjj.
29. Wang, W. & Pfeiffer, T. Securities based decision markets. in *Distributed Artificial Intelligence* (eds. Chen, J., Lang, J., Amato, C. & Zhao, D.) vol. 13170 79–92 (Springer, 2022).
30. Wolfers, J. & Zitzewitz, E. Prediction markets. *J. Econ. Perspect.* **18**, 107–126 (2004).
31. Arrow, K. J. *et al.* The promise of prediction markets. *Science* **320**, 877–878 (2008).
32. Tziralis, G. & Tatsiopoulos, I. Prediction markets: An extended literature review. *J. Predict. Mark.* **1**, 75–91 (2012).
33. Ames, D. L. & Fiske, S. T. Perceived intent motivates people to magnify observed harms. *Proc. Natl. Acad. Sci.* **112**, 3599–3605 (2015).
34. Atir, S. & Ferguson, M. J. How gender determines the way we speak about professionals. *Proc. Natl. Acad. Sci.* **115**, 7278–7283 (2018).
35. Baldwin, M. & Lammers, J. Past-focused environmental comparisons promote proenvironmental outcomes for conservatives. *Proc. Natl. Acad. Sci.* **113**, 14953–14957 (2016).

36. Bear, A., Fortgang, R. G., Bronstein, M. V. & Cannon, T. D. Mistiming of thought and perception predicts delusionality. *Proc. Natl. Acad. Sci.* **114**, 10791–10796 (2017).

37. Boswell, R. G., Sun, W., Suzuki, S. & Kober, H. Training in cognitive strategies reduces eating and improves food choice. *Proc. Natl. Acad. Sci.* **115**, E11238–E11247 (2018).

38. Caruso, E. M., Burns, Z. C. & Converse, B. A. Slow motion increases perceived intent. *Proc. Natl. Acad. Sci.* **113**, 9250–9255 (2016).

39. Casella, A., Kartik, N., Sanchez, L. & Turban, S. Communication in context: Interpreting promises in an experiment on competition and trust. *Proc. Natl. Acad. Sci.* **115**, 933–938 (2018).

40. Chao, M. Demotivating incentives and motivation crowding out in charitable giving. *Proc. Natl. Acad. Sci.* **114**, 7301–7306 (2017).

41. Cheon, B. K. & Hong, Y.-Y. Mere experience of low subjective socioeconomic status stimulates appetite and food intake. *Proc. Natl. Acad. Sci.* **114**, 72–77 (2017).

42. Clarkson, J. J. *et al.* The self-control consequences of political ideology. *Proc. Natl. Acad. Sci.* **112**, 8250–8253 (2015).

43. Cooney, G., Gilbert, D. T. & Wilson, T. D. When fairness matters less than we expect. *Proc. Natl. Acad. Sci.* **113**, 11168–11171 (2016).

44. Côté, S., House, J. & Willer, R. High economic inequality leads higher-income individuals to be less generous. *Proc. Natl. Acad. Sci.* **112**, 15838–15843 (2015).

45. Flesch, T., Balaguer, J., Dekker, R., Nili, H. & Summerfield, C. Comparing continual task learning in minds and machines. *Proc. Natl. Acad. Sci.* **115**, E10313–E10322 (2018).

46. Genschow, O., Rigoni, D. & Brass, M. Belief in free will affects causal attributions when judging others' behavior. *Proc. Natl. Acad. Sci.* **114**, 10071–10076 (2017).

47. Gheorghiu, A. I., Callan, M. J. & Skylark, W. J. Facial appearance affects science communication. *Proc. Natl. Acad. Sci.* **114**, 5970–5975 (2017).

48. Guilbeault, D., Becker, J. & Centola, D. Social learning and partisan bias in the interpretation of climate trends. *Proc. Natl. Acad. Sci.* **115**, 9714–9719 (2018).

49. Halevy, N. & Halali, E. Selfish third parties act as peacemakers by transforming conflicts and promoting cooperation. *Proc. Natl. Acad. Sci.* **112**, 6937–6942 (2015).

50. Handley, I. M., Brown, E. R., Moss-Racusin, C. A. & Smith, J. L. Quality of evidence revealing subtle gender biases in science is in the eye of the beholder. *Proc. Natl. Acad. Sci.* **112**, 13201–13206 (2015).

51. Hoffman, K. M., Trawalter, S., Axt, J. R. & Oliver, M. N. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proc. Natl. Acad. Sci.* **113**, 4296–4301 (2016).

52. Hofstetter, R., Rüppell, R. & John, L. K. Temporary sharing prompts unrestrained disclosures that leave lasting negative impressions. *Proc. Natl. Acad. Sci.* **114**,

11902–11907 (2017).

53. Horne, Z., Powell, D., Hummel, J. E. & Holyoak, K. J. Countering antivaccination attitudes. *Proc. Natl. Acad. Sci.* **112**, 10321–10324 (2015).

54. Isley, S. C., Stern, P. C., Carmichael, S. P., Joseph, K. M. & Arent, D. J. Online purchasing creates opportunities to lower the life cycle carbon footprints of consumer products. *Proc. Natl. Acad. Sci.* **113**, 9780–9785 (2016).

55. Jachimowicz, J. M., Chafik, S., Munrat, S., Prabhu, J. C. & Weber, E. U. Community trust reduces myopic decisions of low-income individuals. *Proc. Natl. Acad. Sci.* **114**, 5401–5406 (2017).

56. John, L. K., Barasz, K. & Norton, M. I. Hiding personal information reveals the worst. *Proc. Natl. Acad. Sci.* **113**, 954–959 (2016).

57. Jordan, J. J., Hoffman, M., Nowak, M. A. & Rand, D. G. Uncalculating cooperation is used to signal trustworthiness. *Proc. Natl. Acad. Sci.* **113**, 8658–8663 (2016).

58. Jun, Y., Meng, R. & Johar, G. V. Perceived social presence reduces fact-checking. *Proc. Natl. Acad. Sci.* **114**, 5976–5981 (2017).

59. KC, R. P., Kunter, M. & Mak, V. The influence of a competition on noncompetitors. *Proc. Natl. Acad. Sci.* **115**, 2716–2721 (2018).

60. Klein, N. & O'Brien, E. People use less information than they think to make up their minds. *Proc. Natl. Acad. Sci.* **115**, 13222–13227 (2018).

61. Kouchaki, M. & Gino, F. Memories of unethical actions become obfuscated over time. *Proc. Natl. Acad. Sci.* **113**, 6166–6171 (2016).

62. Kraus, M. W., Rucker, J. M. & Richeson, J. A. Americans misperceive racial economic equality. *Proc. Natl. Acad. Sci.* **114**, 10324–10331 (2017).

63. McCall, L., Burk, D., Laperrière, M. & Richeson, J. A. Exposure to rising inequality shapes Americans' opportunity beliefs and policy support. *Proc. Natl. Acad. Sci.* **114**, 9593–9598 (2017).

64. Morris, A., MacGlashan, J., Littman, M. L. & Cushman, F. Evolution of flexibility and rigidity in retaliatory punishment. *Proc. Natl. Acad. Sci.* **114**, 10396–10401 (2017).

65. Mummolo, J. Militarization fails to enhance police safety or reduce crime but may harm police reputation. *Proc. Natl. Acad. Sci.* **115**, 9181–9186 (2018).

66. Payne, B. K., Brown-Iannuzzi, J. L. & Hannay, J. W. Economic inequality increases risk taking. *Proc. Natl. Acad. Sci.* **114**, 4643–4648 (2017).

67. Phillips, J. & Cushman, F. Morality constrains the default representation of what is possible. *Proc. Natl. Acad. Sci.* **114**, 4649–4654 (2017).

68. Rai, T. S., Valdesolo, P. & Graham, J. Dehumanization increases instrumental violence, but not moral violence. *Proc. Natl. Acad. Sci.* **114**, 8511–8516 (2017).

69. Reeck, C., Wall, D. & Johnson, E. J. Search predicts and changes patience in intertemporal choice. *Proc. Natl. Acad. Sci.* **114**, 11890–11895 (2017).

70. Schilke, O., Reimann, M. & Cook, K. S. Power decreases trust in social exchange. *Proc. Natl. Acad. Sci.* **112**, 12950–12955 (2015).

71. Stern, C., West, T. V. & Rule, N. O. Conservatives negatively evaluate counterstereotypical people to maintain a sense of certainty. *Proc. Natl. Acad. Sci.* **112**, 15337–15342 (2015).

72. Vacharkulksemsuk, T. *et al.* Dominant, open nonverbal displays are attractive at zero-acquaintance. *Proc. Natl. Acad. Sci.* **113**, 4009–4014 (2016).

73. Williams, K. E. G., Sng, O. & Neuberg, S. L. Ecology-driven stereotypes override race stereotypes. *Proc. Natl. Acad. Sci.* **113**, 310–315 (2016).

74. Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J. & Kievit, R. A. An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* **7**, 632–638 (2012).

75. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl. Acad. Sci.* **115**, 2600–2606 (2018).

76. Benjamin, D. J. *et al.* Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).

77. Simonsohn, U. Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychol. Sci.* **26**, 559–569 (2015).

78. Ly, A., Verhagen, J. & Wagenmakers, E.-J. Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *J. Math. Psychol.* **72**, 19–32 (2016).

79. Ly, A., Etz, A., Marsman, M. & Wagenmakers, E.-J. Replication Bayes factors from evidence updating. *Behav. Res. Methods* **51**, 2498–2508 (2019).

80. Jeffreys, H. *The Theory of Probability*. (Oxford University Press, 1961).

81. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640 (2008).

82. Button, K. S. *et al.* Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).

83. Altmejd, A. *et al.* Predicting the replicability of social science lab experiments. *PLoS One* **14**, e0225826 (2019).

84. Yang, Y., Youyou, W. & Uzzi, B. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proc. Natl. Acad. Sci.* **117**, 10762–10768 (2020).

85. Rajtmajer, S. *et al.* A synthetic prediction market for estimating confidence in published work. *Proc. AAAI Conf. Artif. Intell.* **36**, 13218–13220 (2022).

86. Youyou, W., Yang, Y. & Uzzi, B. A discipline-wide investigation of the replicability of psychology papers over the past two decades. *Proc. Natl. Acad. Sci.* **120**, e2208863120 (2023).

87. Zhou, H. & Fishbach, A. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *J. Pers. Soc. Psychol.* **111**, 493–504 (2016).

88. Chmielewski, M. & Kucker, S. C. An MTurk crisis? Shifts in data quality and the impact on study results. *Soc. Psychol. Personal. Sci.* **11**, 464–473 (2020).

89. Aguinis, H., Villamor, I. & Ramani, R. S. MTurk research: Review and recommendations. *J. Manag.* **47**, 823–837 (2021).

90. Brodeur, A., Cook, N. & Heyes, A. We need to talk about Mechanical Turk: What 22,989 hypothesis tests tell us about publication bias and p-hacking in online experiments. IZA Discussion Paper at (2022).

91. Peer, E., Rothschild, D., Gordon, A., Evernden, Z. & Damer, E. Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods* **54**, 1643–1662 (2022).

92. Webb, M. A. & Tangney, J. P. Too good to be true: Bots and bad data from Mechanical Turk. *Perspect. Psychol. Sci.* 174569162211200 (2022) doi:10/gq7nw4.

93. Douglas, B. D., Ewell, P. J. & Brauer, M. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS One* **18**, e0279720 (2023).

94. Agley, J., Xiao, Y., Nolan, R. & Golzarri-Arroyo, L. Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behav. Res. Methods* **54**, 885–897 (2022).

95. Epstein, R. & Robertson, R. E. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proc. Natl. Acad. Sci.* **112**, E4512–E4521 (2015).

96. Gallo, E. & Yan, C. The effects of reputational and social knowledge on cooperation. *Proc. Natl. Acad. Sci.* **112**, 3647–3652 (2015).

97. Li, V., Michael, E., Balaguer, J., Herce Castañón, S. & Summerfield, C. Gain control explains the effect of distraction in human perceptual, cognitive, and economic decision making. *Proc. Natl. Acad. Sci.* **115**, E8825–E8834 (2018).

98. Hanson, R. Logarithmic market scoring rules for modular combinatorial information aggregation. *J. Predict. Mark.* **1**, 3–15 (2007).

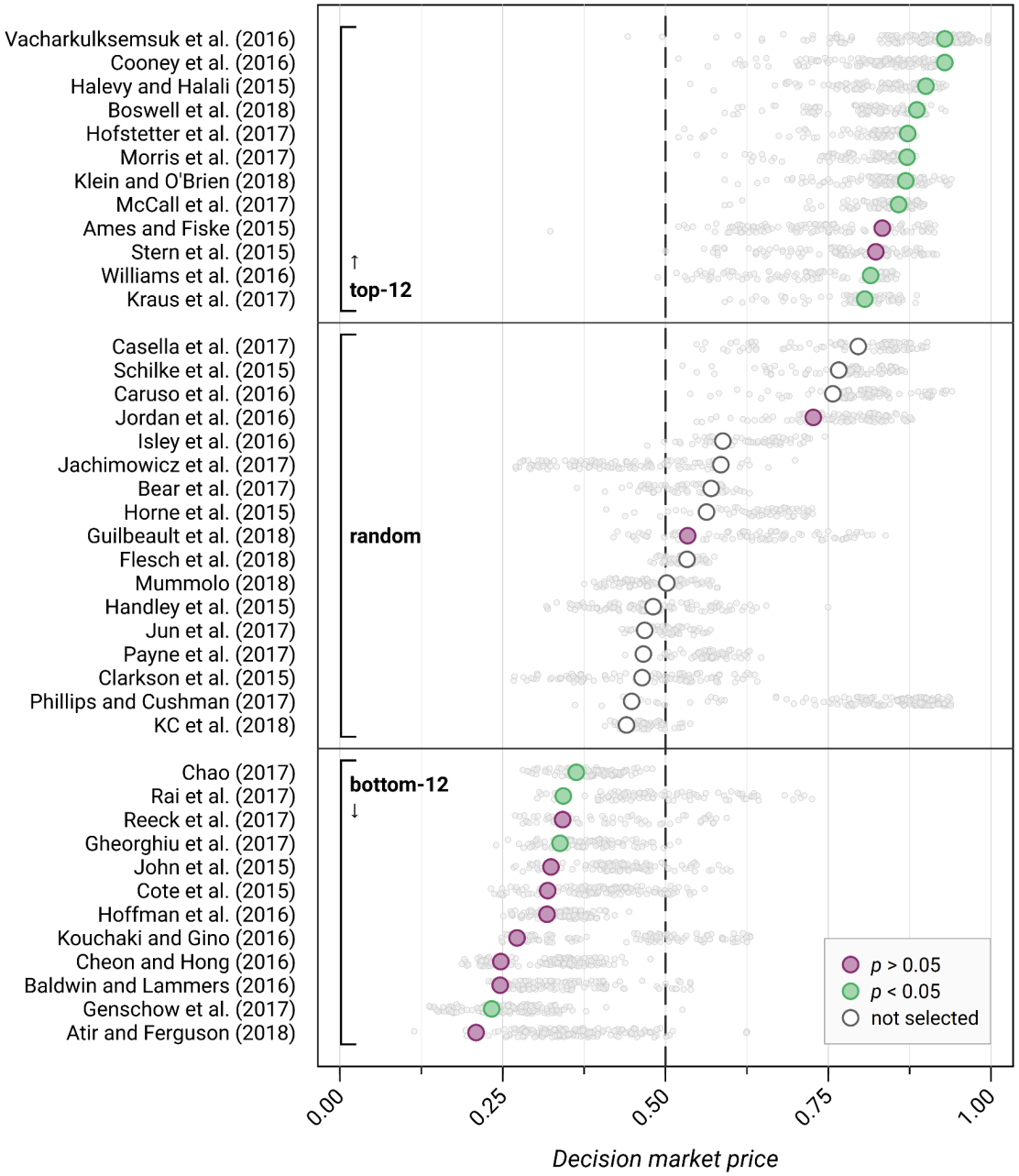

Fig. 1. Decision market prices for the 41 included studies. Plotted are the decision market prices for the 41 *MTurk* social science experiments published in *PNAS* between 2015 and 2018. The small gray dots indicate the market prices after each market transaction; the larger dots indicate the final market price. The studies are ordered based on the final decision market prices, which can be interpreted as the market's probability forecast of successful replication. The 12 studies with the highest decision market prices and the 12 studies with the lowest decision market prices were selected for replication; in addition, two of the remaining 17 studies were selected for replication at random to ensure that the decision market is incentive compatible. The replication outcomes for the statistical significance indicator are also illustrated for the 26 replicated studies. The point-biserial correlation between the decision market prices and the replication outcomes in primary hypothesis 1 is $r = 0.505$ (95% CI [0.146, 0.712], $t(24) = 2.867$, $p = 0.008$; $n = 26$).

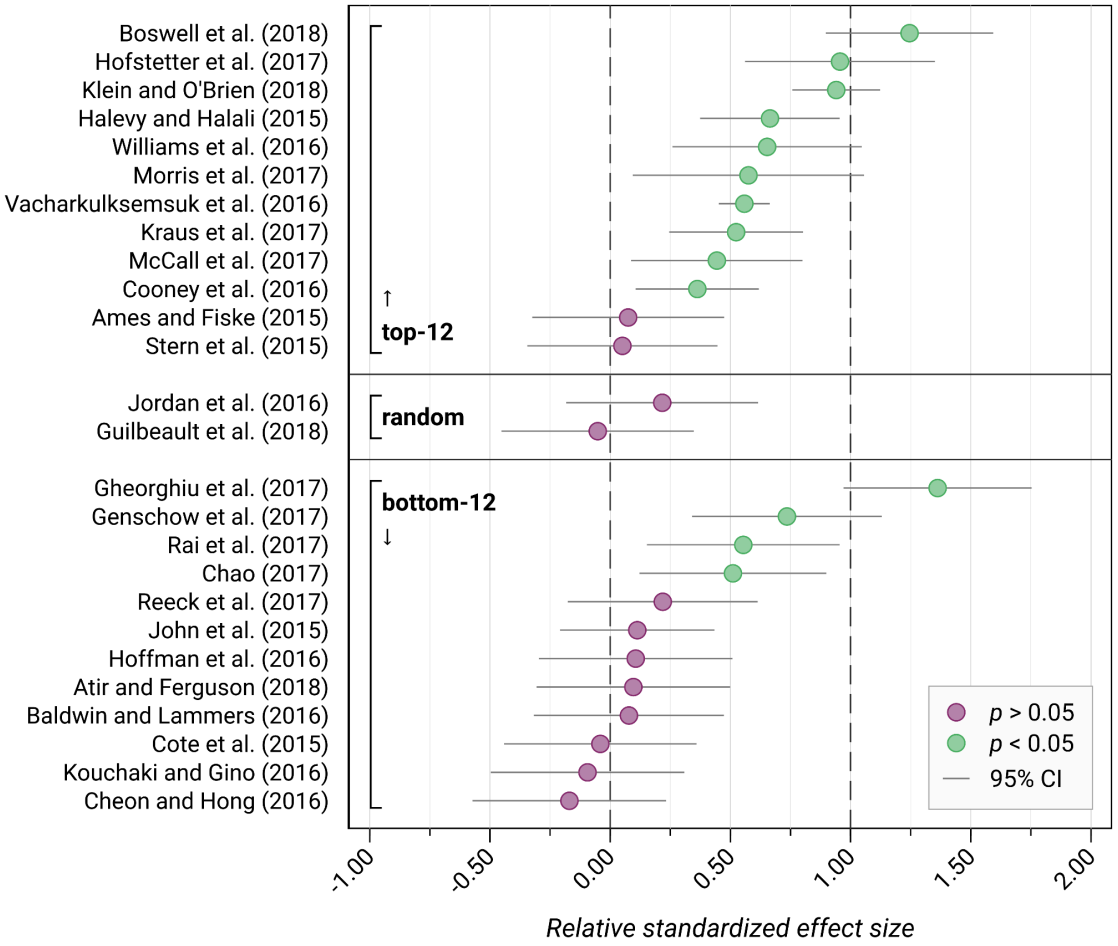

Fig. 2. Replication results. Plotted are the point estimates and the 95% CIs of the 26 replications (standardized to Cohen's *d* units). The standardized effect sizes are normalized such that one equals the original effect size. Studies within each of the three panels (top-12, random, bottom-12) are sorted based on the relative effect size. There is a statistically significant effect ($p < 0.05$) in the same direction as the original study for 14 out of 26 replications (53.8%; 95% CI [33.4%, 73.4%]). For the 12 studies with the highest decision market prices, there is a statistically significant effect ($p < 0.05$) in the same direction as the original study for 10 out of 12 replications (83.3%; 95% CI [51.6%, 97.9%]). For the 12 studies with the lowest decision market prices, there is a statistically significant effect ($p < 0.05$) in the same direction as the original study for 4 out of 12 replications (33.3%; 95% CI [9.9%, 65.1%]). Our secondary hypothesis test provides suggestive evidence that the difference in replication rates between the top-12 and the bottom-12 group is different from zero (Fisher's exact test; $\chi^2(1) = 6.171$ $p = 0.036$; $n = 24$).

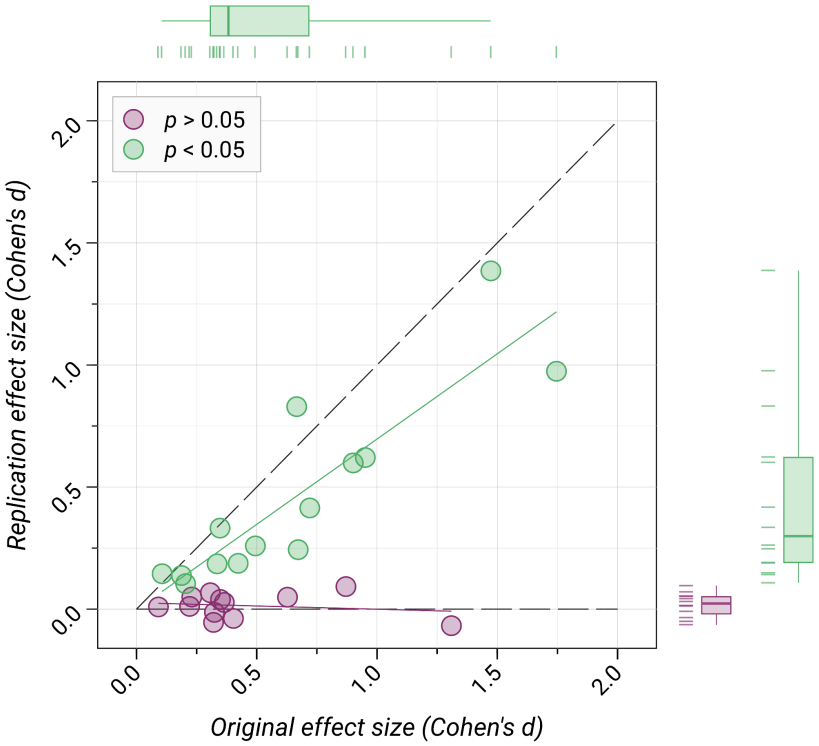

Fig. 3. Relationship between original and replication effect sizes. Plotted are the original and replication effect sizes for each of the 26 replication studies (the effect sizes of both the original and replication studies are standardized to Cohen's *d* units). The mean effect size of the 26 replication studies is 0.253 (*sd* = 0.357) compared to 0.563 (*sd* = 0.426) for the original studies resulting in a relative effect size of 45.0%, confirming our second primary hypothesis (Wilcoxon signed-rank test; $z = 4.203$, $p < 0.001$; $n = 26$). The relative effect size of the 13 replications that have been successfully replicated according to the statistical significance indicator is 69.5%, and the relative effect size of the 13 studies that did not replicate is 3.2%.

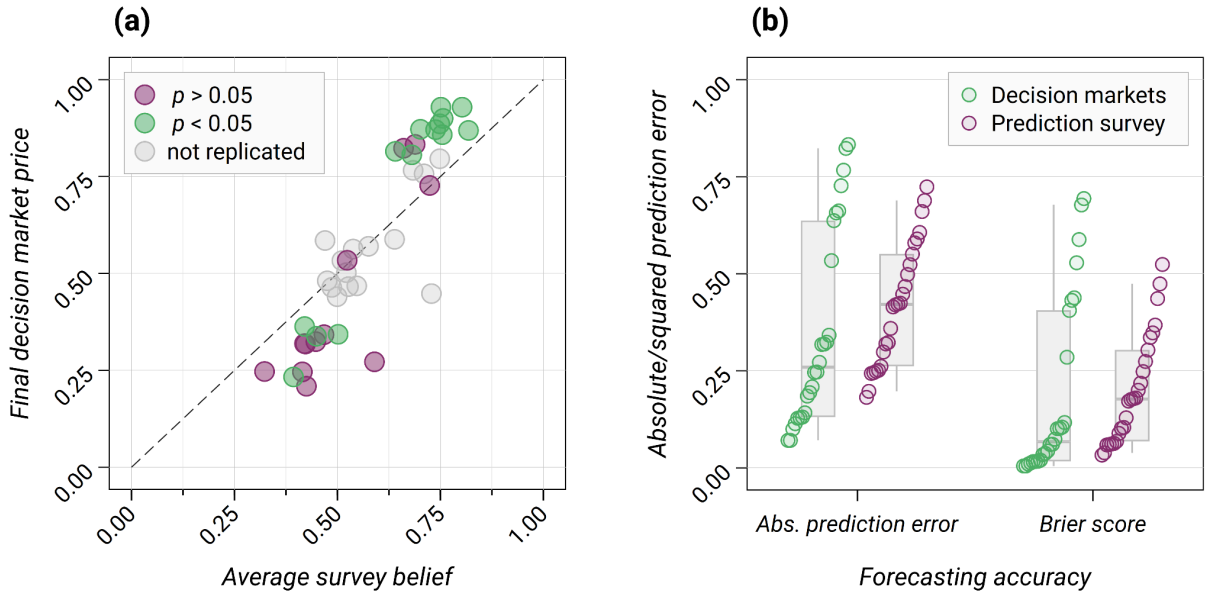

Fig. 4. Relationship between decision market prices and mean survey beliefs and forecasting accuracy. a, Plotted are the decision market prices and the mean survey beliefs about replication for the 41 studies included in the decision market and the survey; the color coding highlights the replication outcomes for the 26 replicated studies. The decision market prices and the mean survey beliefs about replication are highly correlated with a Pearson correlation of $r = 0.899$ (95% CI [0.814, 0.944]; $t(39) = 12.830$, $p < 0.001$; $n = 41$). **b**, Plotted are the absolute prediction errors and the Brier scores (the squared prediction errors) for the decision market and the prediction survey for the 26 replicated studies. There is suggestive evidence of higher prediction accuracy for the decision market in terms of the absolute prediction error (0.353 for the decision markets and 0.421 for the survey; Wilcoxon signed-rank test: $z = 2.172$, $p = 0.030$; $n = 26$), but not in terms of the Brier score (0.188 for the decision markets and 0.202 for the survey; Wilcoxon signed-rank test: $z = 1.181$, $p = 0.238$; $n = 26$))