ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Chandrashekar SP, Viganola D, Dreber A, Johannesson M, Pfeiffer T, Siegel A, Feldman G. 2026 Using prediction markets and forecasting surveys to predict 28 replication outcomes of classic articles in social psychology and judgement and decision making. *R. Soc. Open Sci.* **13**: 250377.
https://doi.org/10.1098/rsos.250377

Received: 24 February 2025
Accepted: 8 September 2025

Subject Category:
Psychology and cognitive neuroscience

Subject Areas:
psychology

Keywords:
meta-science, prediction markets, forecasting, replications, decision making, social psychology

Author for correspondence:
Gilad Feldman
e-mail: giladfel@gmail.com

Supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.8063063.

THE ROYAL SOCIETY PUBLISHING

Using prediction markets and forecasting surveys to predict 28 replication outcomes of classic articles in social psychology and judgement and decision making

Subramanya Prasad Chandrashekar[1], Domenico Viganola[2], Anna Dreber[3,4], Magnus Johannesson[3], Thomas Pfeiffer[6], Adam Siegel[7] and Gilad Feldman[5,8]

[1]Norwegian University of Science and Technology, Trondheim, Norway
[2]World Bank, Washington, DC, USA
[3]Department of Economics, Stockholm School of Economics, Stockholm, Sweden
[4]Department of Economics, and [5]Department of Banking and Finance, University of Innsbruck, Innsbruck, Austria
[6]NZ IAS, Massey University, Auckand, New Zealand
[7]Cultivate Labs, Chicago, IL, USA
[8]Department of Psychology, The University of Hong Kong, Hong Kong, Hong Kong

SPC, 0000-0002-8599-9241; DV, 0000-0003-1545-1990; AD, 0000-0003-3989-9941; MJ, 0000-0001-8759-6393; TP, 0000-0002-0592-577X; GF, 0000-0003-2812-6599

Can researchers predict if classic findings published in the field of social psychology and judgement and decision making replicate? We set up prediction markets and a forecasting survey for predicting replications of 28 experiments of classic well-cited articles. Forecasters predicted if the original results would replicate, where a successful replication was defined as an effect in the same direction as the original and a signal (p-value lower than 0.05). Of the 28 original studies, 16 (57%) met the replication success criteria, compared to a predicted replication rate of 70% in the prediction markets and 65% replication rate in the survey. We concluded only suggestive evidence for associations of replication outcomes with prediction market prices (r = 0.43, 95% CI [0.07, 0.69]) and average survey beliefs (r = 0.26, 95% CI [–0.12, 0.58]).

The prediction market effects were similar to observed effects in previous prediction market studies and suggest that prediction markets can to some extent predict replication outcomes, yet predictions are far from perfect and conducting replications is much more informative about the credibility of published findings. Data and code are available at https://doi.org/10.17605/OSF.IO/2KMH7.

1. Predictions of replicability

Which research results can we trust? This question is at the core of the scientific endeavour, and one way to address this question is to conduct replications of published research results. By replication we mean testing the same hypothesis as in the original study with newly collected data. We classify a replication as a 'direct replication' if a very similar research design and analysis is used as in the original study [1,2]. Recently a number of systematic replication projects have been conducted in the social sciences, carrying out direct replications of systematically selected original studies (e.g. [3–11]). Combining the results from these systematic replication projects in social psychology suggests a replication rate of about 30–50% [12] of the replications with an effect in the hypothesized direction that detected a signal at the set alpha level and with replication effect sizes that were about half of the effects reported in the original studies. These led to a series of meta-science projects to try and identify factors that would help explain these results and identify ways to improve research practices so that future studies would meet higher replicability rates. Some of the meta-science findings suggested that, for example, some researchers employ 'questionable research practices' in which researchers, either consciously or subconsciously, take advantage of researcher degrees of freedom in research design, methodology, and analyses to increase the chances of obtaining statistically significant results [13–16], which are more likely to be published in scientific journals given a 'publication bias' in journals' very strong preference for publishing positive (non-null) findings [17–19]. Another example are findings on the prevalence of underpowered studies [20,21] and that even experienced researchers tend to draw unwarranted conclusions and assuming high replicability of studies conducted based on small samples [22,23]. These have led to a call for a science reform to improve scientific practices and increase rigour and transparency, adopt solutions such as pre-registration of analysis plans [24], Registered Reports [25,26], replications [27,28] and large team science [29].

Is it possible to accurately predict replication outcomes? This question is important for several reasons. First, it gives an indication of the added value of conducting replications. If researchers can accurately predict which original results replicate well, then it might be possible to reduce the vast resources and money invested in conducting replications by instead surveying researchers and aggregating predictions about replication results. Second, if predictions are imperfect but are still associated with replication outcomes, then researcher replication predictions might be used to prioritize replications that might offer the biggest benefits per dollar spent, such as in prioritizing the replications of studies that scholars are less sure about or have less agreement about [30]. Third, researcher predictions of replicability can be used as another source of information that, when combined with actual replication results, may help the academic community better estimate prior and posterior probabilities of the tested hypotheses being true [31]. Finally, prediction markets can help indicate if researchers on average over- or underestimate the credibility of published findings, so that researchers may over time learn to improve to become better calibrated in their expectations of replicability when assessing old findings or in their evaluations of the likelihood of the replicability of new findings.

2. Prediction markets and forecasting surveys

There are different tools for eliciting researcher predictions about the replicability of scientific results. One tool is prediction markets, which have been extensively used to make predictions in other domains such as predicting election outcomes [32–34]. In these markets, participants typically trade contracts on binary outcomes such as whether a study will replicate or not according to a predefined criterion, where contracts are worth some positive monetary amount if the study replicates and 0 if it does not. The price of such contracts can be interpreted as the probability that the study will replicate successfully according to the market. Several of the systematic replication projects mentioned

above have been accompanied by prediction markets to test to what extent prediction markets can predict replication outcomes [3,4,31,35]. Gordon *et al*. [36] pooled the data from these studies and found a correlation of 0.58 between prediction market prices and replication outcomes, and their findings suggest that forecasters on average are overoptimistic about the likelihood of replication with an average overestimation of the replication rate by 14 percentage units. Taken together these studies suggest that prediction markets can be useful tools to aggregate information about researchers' beliefs about replication results; but the predictions are far from perfect.

Another tool to assess whether replication results are predictable is to use forecasting surveys. In the prediction market studies referred to above, participants were also asked to predict the replication outcomes in a survey that preceded the prediction markets. In the survey, forecasters were asked to predict the probability that the study would replicate, defined as an effect in the original direction and a *p*-value below 0.05. Gordon *et al*. [36] pooled the survey prediction data from these studies as well and found support for prediction markets providing more accurate predictions than the average survey predictions, in terms of the absolute prediction error. Yet, the magnitude of this difference was relatively small and the correlation of 0.56 between survey predictions and replication outcomes was almost as high as for the prediction markets. The overestimation of the replication rate in the prediction surveys of 12 percentage units was also similar to the overestimation in the prediction markets. There is now also substantial work using forecasting surveys to predict new outcomes in addition to replication outcomes (e.g. [37,38]), and these studies also suggest a positive association between predictions and outcomes.

3. Current study: CORE team replications of classics in social psychology and decision making

In this study, we contribute to this literature by using prediction markets and forecasting surveys to predict the replication results of 28 replications of original results published in the field of social psychology and judgement and decision making (JDM). The 28 replication studies were part of a larger initiative by the Collaborative Open-science and meta REsearch (CORE) team [39]—an international team of early-career researchers with over 400 students from the University of Hong Kong that engaged in replication efforts aiming to practice and promote best-practices open-science and meta research. The team completed over 120 direct replications, and we aimed to predict a total of 28 of these started in the years 2019 (12 studies) and 2020 (16 studies). The prediction markets covered all studies planned by the CORE team for the years 2019 and 2020 at the beginning of the academic year.

We report an association between replication outcomes and prediction market prices with a correlation of $r = 0.43$ [0.07, 0.69] and an association between average survey beliefs and replication outcomes was weaker with a correlation of $r = 0.26$ [−0.12, 0.58]. On average forecasters were somewhat over-optimistic with a predicted replication rate of 70% in the prediction markets and 65% in the prediction survey, compared to the observed replication rate of 57%. Our results were overall in line with previous studies although the association between forecasts and replications is slightly weaker than in previous studies, especially for the survey predictions. The results suggest that prediction markets may have some ability to predict replications, but that the predictions are imperfect and should be thought of more of a complement than substitute for replications. We note that given the lack of preregistration in our reported analyses and the small number of predicted studies our results should be interpreted with caution and preferably in aggregate together with other similar studies.

4. Method

4.1. Open science

We provide all data and code at: https://doi.org/10.17605/OSF.IO/2KMH7. We report and discuss the sample size (number of replications included and number of forecasters), and report all measures and exclusions.

4.2. Replications

The 28 replication studies were conducted by the CORE team in the years 2019 and 2020. Undergraduate and taught masters (MSc) students conducted replications of classics in social psychology and JDM in one-semester courses and one-year guided thesis work at the University of Hong Kong. In 15 of the 28 replication projects, early-career researchers then joined to take the lead on the projects, they verified the analyses and findings and prepared preprints for public dissemination and submission to scientific journals for publication.

The replication projects were chosen by the team coordinator based on his assessment of factors such as impact on subsequent literature in the field (e.g. citation counts). For example, as of May 2024, the Google Scholar citation count for the 28 replication targets ranged from 36 to 8847 with a median of 1039. The aim of the replication projects was to try and replicate the original study with a methodology as close as possible to the original study.

The replication studies did not only involve testing key hypotheses but also corollary results. In many instances, they also involved testing theoretical extensions. However, from the perspective of replication market predictions, we focused on one key original result being replicated within each target study (which is predicted in the prediction market and the survey). The criterion for a successful replication of a result was that the replication had to detect a signal at an alpha level of 5% with the effect being in the same direction as in the original study. In some studies, the replication involved more than one original result, and we then provided additional information to forecasters on how many of these results need to be detected for it to count as a successful replication. We provide additional information in the electronic supplementary material section about the key hypotheses tested in the replications and information about the original studies and the replication studies. We summarize the 28 replication studies in electronic supplementary material, tables S1 and S2.

4.3. Prediction markets and forecasting survey

Participants were informed that the replications would have at least 95% power to detect the original effect size, that the studies would be performed on either Amazon Mechanical Turk using CloudResearch [40] or Prolific [41], with either a United States sample or a British sample, and that the replications would be conducted as a pre-registered replication where the pre-registration plan is first uploaded to the Open Science Framework and frozen with a public timestamp. The key predictions in all the chosen target studies for replication were statistically significant ($p < 0.05$).

Participants of the forecasting surveys and prediction markets were recruited through announcements on mailing lists (e.g. Society for Judgement and Decision Making) as well as social media networks (e.g. the corresponding author's Twitter/X feed). Interested participants signed up to take part in the study via a short survey. The survey provided prospective study participants with information regarding the background, eligibility, pay and schedule of the study. The only exclusion criterion for participating in the study was being part of the CORE team and/or involved in running these replications. In the sign-up survey, we asked potential participants whether they were members of the replication team or had any insider information about the replication process, and then asked for each participant's name, affiliation, position and country.

Participants had access to the target articles, the replication pre-registrations and a short summary created by the authors to help participants make an informed decision.

Before participating in the prediction markets, participants completed a forecasting survey. For the first round of the forecasting survey and prediction market, 136 participants signed up, one participant was excluded from participating as they were a member of the replication team, 88 completed the forecasting survey and 64 participated in the prediction markets. For the second survey and prediction markets, 88 participants signed up, 0 participants were excluded from participating, 60 completed the forecasting survey and 50 participated in the prediction markets. That some participants drop out between signing up and completing the forecasting survey, and between completing the forecasting survey and participating in the prediction markets is in line with similar previous prediction market studies [3,4,31,35]. As in previous prediction market studies [3,4,31,35], we excluded participants who participated in the survey but did not make at least one trade in the prediction markets. For the second survey and prediction market, we could not match three of the 50 market participants to a completed survey, and we therefore included the 47 market participants that could be matched to 47 completed surveys in the analyses based on the survey data. Our results for the survey data are therefore based on 64 participants for the first survey and prediction markets and 47 participants for the second survey

and prediction markets. In a robustness test reported in the electronic supplementary material, we also provided results for the prediction survey including all the survey responses (88 for the first survey and 60 for the second survey); the results were very similar to those we report below with the exclusions.

In the forecasting surveys and prediction markets, participants were asked to predict or bet on a binary outcome for each study—whether the replication study's effect would be in the same direction as the original one and statistically significant ($p < 0.05$ in a two-sided test); the round 1 replication studies involved two independent data samples as part of the replication and these were pooled in the replication tests for these studies. In some replications, as detailed in electronic supplementary material, table S1, the replication tests which were predicted were based on more than one test, and it was in these cases specified how many of these tested that needed to have a p-value below 0.05 and an effect in the original direction to count as a successful replication. This was the case for eight replication studies, and the predictions for those eight studies were more complicated than in previous prediction market studies of replications. The predictions focused on the main replication tests, yet some replications also involved additional original results that were replicated (but not predicted), and this means that the conclusions about whether the study replicated or not in some cases differ between the predicted results and the overall replication results. Participants were provided with information about the original study as well as information about the replications; see the electronic supplementary material for the information provided to participants about each of the 28 replicated studies.

In the survey, participants were asked to indicate the probability of each study replicating, whereas in the markets participants were given the opportunity to trade contracts on whether this outcome would occur.

The prediction markets were subsidized. Participants received USD50, corresponding to 10 000 points, on their prediction market account to trade with. For each hypothesis, participants could see the current market prediction for the probability of a successful replication outcome. This probability is equivalent to the price at which contracts were traded, and could thus range from 0.00 to 100.00. Bets could be placed in two different ways: participants could either directly place their bets in terms of points invested, where participants first specified whether they took a long or short position (where a long (short) position implies that participants believe the market price will increase (decrease) or that the true probability that the hypothesis will be supported is higher (lower) than the one identified by the current price), or by entering their beliefs about the probability for support of the hypothesis (ranging from 0 to 100) with the system then making a 'suggested trade' that moved the current market price in the direction of the forecast, though typically not all the way so that the subjective probability and market price would be the same. Participants could repeatedly trade on contracts during the two-week period of the markets.

The first round of markets was conducted between 10 December and 23 December 2019, and the second round was conducted from 1 September to 15 September 2020. The prediction markets were run in collaboration with Cultivate Labs (https://www.cultivatelabs.com/).

Payments were settled in 2020 and 2021 according to the replication results at the time of payments, with payments thus depending on how well the predictions matched the actual results. Any points that were not traded were not included in the payment calculations. Payments were paid out as Amazon gift cards.

4.4. Statistical tests and power

We had no pre-registered analysis plan for the analyses conducted in this paper, which means that the results should be interpreted cautiously and, given the small sample, we focused on the effects rather than the use of p-values in null hypothesis significance testing (NHST). In our analyses, we followed the analyses conducted in previous similar prediction market studies [3,4,31,35]. In all NHST tests below, following Benjamin *et al*. [42], we consider results with p-values below 0.005 to be statistically significant evidence, and results with p-values below 0.05 are interpreted as suggestive evidence (both two-sided). Our primary test was whether the prediction market prices are correlated with a successful replication (yes/no). We ran a sensitivity analysis which showed that with a sample of 28 aiming for 80%, we could only detect an effect of $r = 0.57$ with an alpha set to 0.005 and $r = 0.43$ with an alpha set to 0.05. Gordon *et al*. [36] pooled the data from several prediction market projects predicting replications and reported a correlation between prediction market prices and successful replication of 0.58 in the pooled data.

5. Results

In the first round of the prediction markets, the average number of traders that traded on each replication was 48 (range: 43–56) and the average number of transactions per replication was 95 (range: 67– 141). In the second round, the average number of traders that traded on each replication was 40 (range: 37–45) and the average number of transactions per replication was 68 (range: 50–100). The predicted probability of replication, i.e. the final market prices, ranged from 0.35 to 0.95 in the 28 prediction markets ($m = 0.70$; s.d. = 0.17), whereas the average survey predictions ranged from 0.45 to 0.82 ($m = 0.64$; s.d. = 0.92). We provide descriptive statistics for each market in electronic supplementary material, table S3.

In figure 1, we summarize the prediction market prices and average survey beliefs of the 28 replication studies ordered from the lowest to the highest prediction market price; and with an indicator for whether the study was replicated successfully or not.

One measure of the predictive ability of markets and surveys reported in previous prediction market studies was to estimate the fraction of correct predictions; where a predicted probability over 0.5 is interpreted as predicting replication success and a predicted probability below 0.5 is interpreted as predicting replication failure. As can be seen in figure 1, the prediction markets predicted 18 of the 28 (64%) replications correctly. The average forecasting survey also predicted 18 of the 28 (64%) replications correctly. In the pooled data in Gordon *et al*. [36], these fractions are 73% for prediction markets and 66% for the average survey predictions.

The point biserial correlation coefficient between prediction market prices and replication outcomes was 0.43 (95% CI [0.07, 0.69], $p = 0.022$, d.f. = 26), and we therefore concluded only suggestive evidence of a positive correlation between prediction market prices and replication outcomes. The point-biserial correlation between the average survey beliefs and replication outcomes was 0.26 (95% CI [−0.12, 0.58], $p = 0.179$, d.f. = 26), and we therefore concluded failing to reject the null hypothesis. To allow for a comparison, we note that in the pooled data from previous prediction market studies, Gordon *et al*. [36] reported a correlation of 0.58 between prediction market prices and replication outcomes and a correlation of 0.56 between average survey forecasts and replication outcomes.

The two prediction markets we ran were conducted at two different time points, and we therefore also examined the association between predictions and replication outcomes separately in the two markets. The association between prediction market prices and replication outcomes in the first set of prediction markets was 0.08 (95% CI [−0.52, 0.63], $p = 0.801$, d.f. = 10), and in the second set was 0.56 (95% CI [0.09, 0.83], $p = 0.024$, d.f. = 14; z-test differences between the two: $z = 1.27$, $p = 0.203$). The difference in the point estimates of the correlation between the two sets of markets may be due to random variation between the two sets of markets due to the small sample size in each set (although we are not well-powered to detect an actual difference between the two sets of markets and this test is thus not very informative).

The corresponding correlations between the average survey forecasting and the replication outcomes was in the first round −0.02 (95% CI [−0.58, 0.56], $p = 0.961$, d.f. = 10), and in the second round 0.25 (95% CI [−0.28, 0.66], $p = 0.346$, d.f. = 14; z-test differences between the two: $z = 0.64$, $p = 0.525$). We caution against overinterpretation of the null given the low power to detect differences.

We plot the relationship between the prediction market prices and average survey beliefs in figure 2. The two prediction measures were closely related with a Pearson correlation of 0.82 (95% CI [0.64, 0.91], $p < 0.001$, d.f. = 26), and this high correlation was also in line with previous studies [3,4,31,35].

We also tested if the average absolute prediction error and the average squared prediction error (the Brier score) differed between the prediction market and survey predictions (the absolute prediction error is defined as the absolute difference between the predicted probability of a successful replication and the observed outcome, where the observed outcome = 1 for a successful replication and = 0 for a failed replication). The average (squared) prediction error was 0.40 (0.22) for the prediction markets, and 0.46 (0.24) for the average survey beliefs. We find that there is suggestive evidence of a difference in prediction error between the two methods (paired t-test of a difference in average prediction error: diff = −0.05, 95% CI [−0.10, −0.01], $t(27) = -2.72$, $p = 0.011$). However, no evidence of a difference is observed for squared prediction errors: diff = −0.02, 95% CI [−0.06, 0.02], $t(27) = -0.83$, $p = 0.415$.

Previous work on prediction markets with replication outcomes found indication for an overestimation [36], and our results point to an observed replication rate of 0.57, with an average prediction of 0.70 for the prediction markets and an average prediction of 0.65 for the average forecasting survey. We conducted a paired t-test for the prediction markets and found diff = 0.13, 95% CI [−0.05, −0.31], $t(27)$

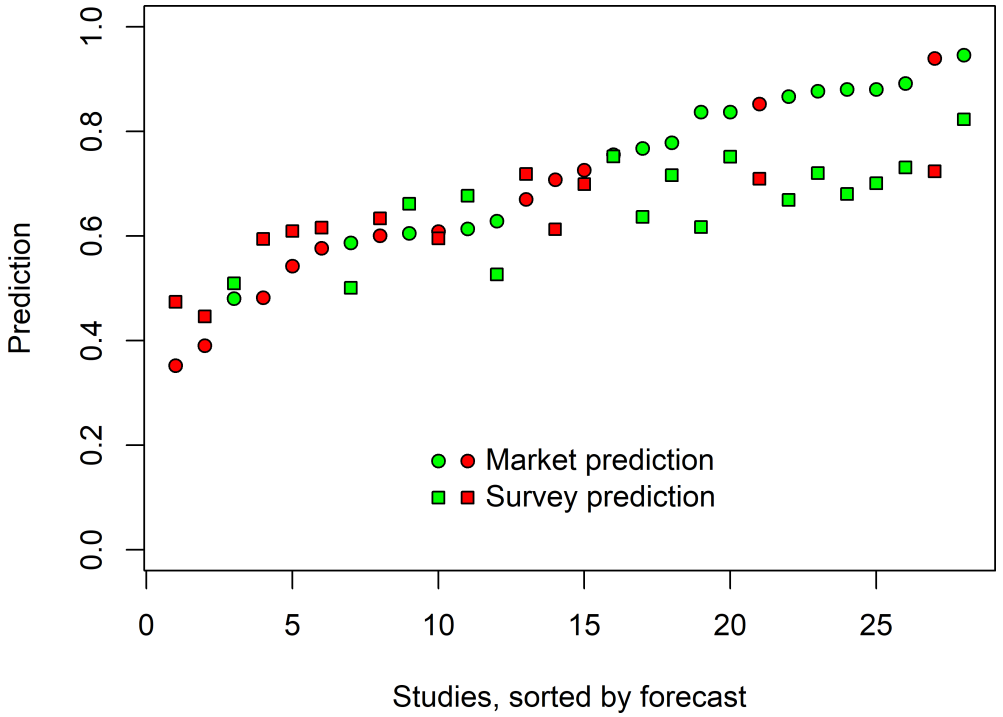

Figure 1. Prediction market and forecasting survey predictions and outcomes. Green symbols represent a successful replication ($p < 0.05$); red symbols represent an unsuccessful replication ($p > 0.05$).

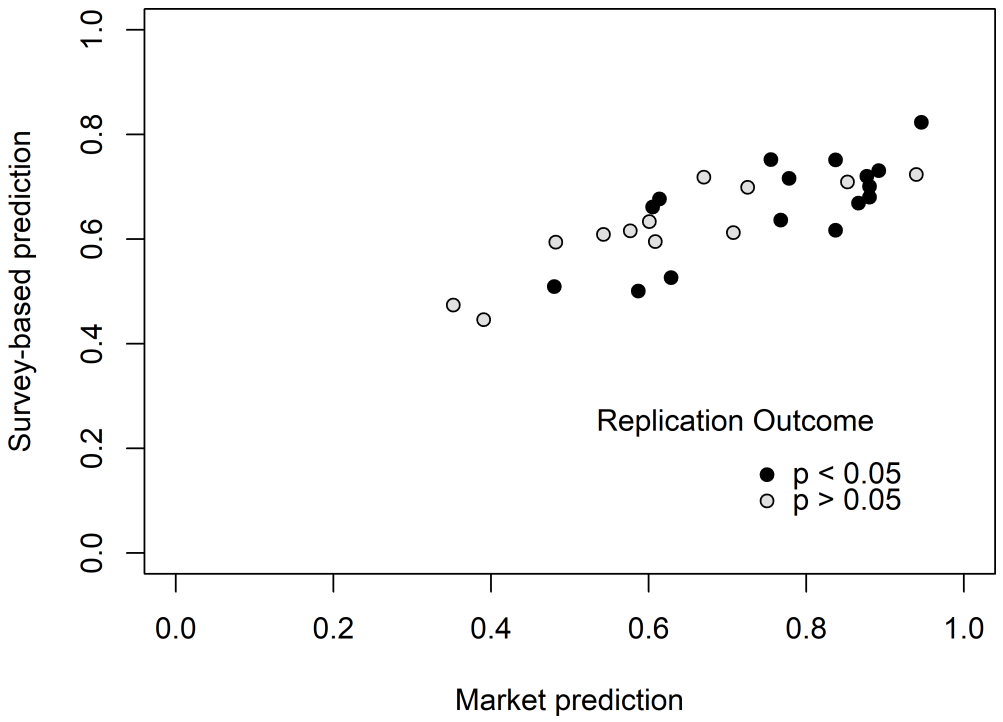

Figure 2. Market prediction versus survey-based prediction.

= 1.52, $p = 0.140$, and for the paired t-test for the surveys, we found diff = 0.08, 95% CI [−0.11, 0.26], $t(27) = 0.82$, $p = 0.422$. However, we caution against overinterpreting these tests given the low power to detect such differences (minimum detectable effect size for 80% statistical power in the test of whether prediction markets overestimate the replication outcomes is 25 percentage units for an alpha of 0.05 and 32 percentage units for an alpha of 0.005).

6. Discussion

We ran prediction markets and forecasting surveys on 28 CORE team replications in two time points, and reported an association between replication outcomes and prediction market prices of 0.43, with a weaker association for forecasting survey predictions of 0.26, and prediction market prices and forecasting survey predictions with an association of 0.82.

These results are largely consistent with the results of previous prediction market studies, albeit with weaker associations (0.43 and 0.26, compared to 0.58 and 0.56 reported in Gordon *et al*. [36]).

In our analyses, we followed the same strategy as in previous prediction market studies [3,4,31,35]. Combined with the results of previous studies, there is evidence that prediction markets have some ability to predict replications, yet that the predictive ability is limited and does not provide a strong substitute for replication studies.

6.1. Prediction markets and forecasting surveys

The choice between prediction markets and surveys is not obvious. In the pooled data in Gordon *et al*. [36], as well as in this study, prediction markets seem to have somewhat better prediction accuracy, yet it is not clear whether this marginal improvement in accuracy is worth the additional costs of collecting prediction market data compared to surveys. One caveat about the survey predictions in this study and previous prediction market studies is that the surveys preceded the prediction markets, and it cannot be ruled out that this improves the survey predictions compared to using prediction surveys without prediction markets. We recommend that future research would aim to test that by conducting a meta study where the order of the forecasting survey and the prediction markets is randomized.

6.2. Limitations, recommendations and future directions

The indicator for a successful replication we used for predictions in the prediction markets and forecasting surveys was the statistical significance indicator of replication, with a successful replication defined as an effect in the original direction and a *p*-value lower than 0.05. One difference compared to previous prediction market studies of replications, is that for eight of the 28 studies the prediction involved multiple original results with a specification of how many that needed to be significant in the original direction with a *p*-value lower than 0.05 to count as a successful replication (see electronic supplementary material, table S1, for further information). The prediction scenario for these eight studies therefore involves a more complicated prediction scenario compared to previous prediction market studies on replications, and it cannot be ruled out that this affects the accuracy of the predictions. It should also be noted that the replication studies *per se* in some cases involved additional original results not part of the predictions in the prediction markets. For some studies, the overall conclusion regarding replication success, therefore, differs from the conclusion about whether the original study replicated or not based on the results predicted in the prediction markets (see electronic supplementary material for references to the reports detailing the overall conclusions about replicability for the 15 replication studies published as preprints or peer reviewed papers). It should also be noted that additional replication indicators could be used to measure replicability, such as the relative effect size indicator measuring the effect sizes of the replications relative to the original effect sizes. The relative effect size replication indicator is a continuous indicator of replicability, which we recommend to always report alongside the statistical significance indicator in replication studies. However, in this replication study, this is complicated by missing information about the original effect sizes for some of the replication studies included in the prediction markets; see electronic supplementary material, table S1, for details.

There are also other potential prediction methods. One possibility would be to predict the replication probability based on the original *p*-value. Based on pooled data from the replication studies included in previous prediction market studies, Gordon *et al*. [36] reported a correlation between the original *p*-value and the replication outcome of 0.46, based only on categorizing the original *p*-values as above or below 0.005 with a replication rate of 74% for original results with *p*-values less than 0.005 and a replication rate of 28% of original results reported as statistically significant with a *p*-value over 0.005. The correlation between the original *p*-values and the replication outcomes in this study is not straightforward to estimate as some of the predictions involved multiple tests and thereby multiple

original *p*-values, and for some of the original studies, we lack information about the exact original *p*-values or whether they were above or below 0.005. Even if the prediction accuracy is somewhat lower using original *p*-values than using prediction markets or prediction surveys, it is a more practical alternative as no additional data collection is needed. Another option is using machine learning models to predict replication outcomes. A number of prediction models based on machine learning have been published in recent years [43–46], but it is not yet clear if these models can outperform predictions from prediction markets and surveys (they should be able to predict at least as well as based on original *p*-values by using that information).

Our sample included only 28 replications, which severely limited our statistical power to detect associations and differences using NHST. Another limitation is that we did not pre-register any of our analyses. Given the two limitations, we consider the findings exploratory and suggest caution in overinterpreting null findings. We recommend focusing on the effect sizes and examining our results in combination with other prediction markets and forecasting surveys to draw larger meta-scientific insights on the broader literature.

We did not collect full information regarding the participants' research background, yet we note that the research understanding and the sustained engagement required by the study was a strong barrier to participation by individuals without relevant expertise. Future research may improve on our design to collect more background information about the participants and examine how these might be associated with prediction accuracy.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. We provide all data and code on OSF [47]. We report and discuss the sample size (number of replications included and number of forecasters), and report all measures and exclusions.

Electronic supplementary material is available online [48].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. S.P.C.: formal analysis, investigation, methodology, writing—original draft; D.V.: conceptualization, data curation, methodology, writing—review and editing; A.D.: conceptualization, funding acquisition, methodology, project administration, writing—original draft, writing—review and editing; M.J.: conceptualization, methodology, writing—original draft, writing—review and editing; T.P.: conceptualization, formal analysis, methodology, writing—review and editing; A.S.: data curation, methodology, writing—review and editing; G.F.: conceptualization, investigation, project administration, resources, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. Some of our authors have published several other articles with similar methods on prediction markets. The first and corresponding authors are members of the CORE team that conducted the replications that were the target of the prediction market, and the corresponding author is the coordinator of that team. One author is a member of Cultivate Labs, which is a platform for conducting prediction markets.

Funding. We are grateful for funding from the Jan Wallander and Tom Hedelius Foundation (P23-0098), the Knut and Alice Wallenberg Foundation, the Marianne and Marcus Wallenberg Foundation, and the Riksbankens Jubileumsfond (P21-0168). Cultivate Labs employs A.S. and provided the online market interface used in the study. The market interface is commercial software. The replication projects that were the target for the prediction markets were supported by the Teaching Development Grant from the University of Hong Kong awarded to G.F. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

References

1. Dreber A, Johannesson M. 2024 A framework for evaluating reproducibility and replicability in economics. *Econ. Inq.* **63**, 338–356. (doi:10.1111/ecin.13244)
2. LeBel EP, McCarthy RJ, Earp BD, Elson M, Vanpaemel W. 2018 A unified framework to quantify the credibility of scientific findings. *Adv. Methods Pract. Psychol. Sci.* **1**, 389–402. (doi:10.1177/2515245918787489)
3. Camerer CF *et al*. 2016 Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436. (doi:10.1126/science.aaf0918)
4. Camerer CF *et al*. 2018 Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644. (doi:10.1038/s41562-018-0399-z)
5. Ebersole CR *et al*. 2016 Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82. (doi:10.1016/j.jesp.2015.10.012)
6. Ebersole CR *et al*. 2020 Many Labs 5: testing pre-data-collection peer review as an intervention to increase replicability. *Adv. Methods Pract. Psychol. Sci.* **3**, 309–331. (doi:10.1177/2515245920958687)

7. Holzmeister F *et al*. 2025 Examining the replicability of online experiments selected by a decision market. *Nat. Hum. Behav*. **9**, 316–330. (doi:10.1038/s41562-024-02062-9)
8. Klein RA. 2014 Investigating variation in replicability a 'many labs' replication project. *Soc. Psychol.* **45**, 142–152. (doi:10.1027/1864-9335/a000178)
9. Klein RA *et al*. 2018 Many Labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490. (doi:10.1177/2515245918810225)
10. Klein RA *et al*. 2022 Many Labs 4: failure to replicate mortality salience effect with and without original author involvement. *Collabra* **8**, 35271. (doi:10.1525/collabra.35271)
11. Open Science Collaboration. 2015 Estimating the reproducibility of psychological science. *Science* **349**, aac4716. (doi:10.1126/science.aac4716)
12. Feldman G. 2022 'Endorsing open science challenges and getting started', talk by Gilad Feldman to the University of Buenos Aires. OSF. (doi:10.17605/OSF.IO/B9SGR)
13. Gelman A, Loken E. 2014 The statistical crisis in science. *Am. Sci*. **102**, 460–465. (doi:10.1515/9781400873371-028)
14. John LK, Loewenstein G, Prelec D. 2012 Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci*. **23**, 524–532. (doi:10.1177/0956797611430953)
15. Nelson LD, Simmons J, Simonsohn U. 2018 Psychology's renaissance. *Annu. Rev. Psychol*. **69**, 511–534. (doi:10.1146/annurev-psych-122216-011836)
16. Simmons JP, Nelson LD, Simonsohn U. 2011 False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci*. **22**, 1359–1366. (doi:10.1177/0956797611417632)
17. Chambers C. 2017 *The seven deadly sins of psychology: a manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.
18. Ferguson CJ, Heene M. 2012 A vast graveyard of undead theories: publication bias and psychological science's aversion to the null. *Perspect. Psychol. Sci.* **7**, 555–561. (doi:10.1177/1745691612459059)
19. Hardwicke TE, Serghiou S, Janiaud P, Danchev V, Crüwell S, Goodman SN, Ioannidis JPA. 2020 Calibrating the scientific ecosystem through meta-research. *Annu. Rev. Stat. Appl.* **7**, 11–37. (doi:10.1146/annurev-statistics-031219-041104)
20. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR. 2013 Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci*. **14**, 365–376. (doi:10.1038/nrn3475)
21. Ioannidis JPA, Stanley TD, Doucouliagos H. 2017 The power of bias in economics research. *Econ. J*. **127**, F236–F265. (doi:10.1111/ecoj.12461)
22. Hong C, Feldman G. 2024 Revisiting the 'Belief in the law of small numbers': conceptual replication and extensions Registered Report of problems reviewed in Tversky and Kahneman (1971). OSF. (doi:10.17605/OSF.IO/MNS7J)
23. Tversky A, Kahneman D. 1971 Belief in the law of small numbers. *Psychol. Bull*. **76**, 105. (doi:10.1037/h0031322)
24. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. 2018 The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606. (doi:10.1073/pnas.1708274114)
25. Nosek BA, Lakens D. 2014 Editorial. Registered reports: a method to increase the credibility of published results. *Soc. Psychol*. **45**, 137–141. (doi:10.1027/1864-9335/a000192)
26. Scheel AM, Schijen M, Lakens D. 2021 An excess of positive results: comparing the standard psychology literature with registered reports. *Adv. Methods Pract. Psychol. Sci*. **4**, 251524592110074. (doi:10.1177/25152459211007467)
27. Nosek BA *et al*. 2022 Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol*. **73**, 719–748. (doi:10.1146/annurev-psych-020821-114157)
28. Zwaan RA, Etz A, Lucas RE, Donnellan MB. 2018 Making replication mainstream. *Behav. Brain Sci*. **41**, e120. (doi:10.1017/S0140525X17001972)
29. Forscher PS, Wagenmakers EJ, Coles NA, Silan MA, Dutra N, Basnight-Brown D, IJzerman H. 2023 The benefits, barriers, and risks of big-team science. *Perspect. Psychol. Sci.* **18**, 607–623. (doi:10.1177/17456916221082970)
30. Isager PM *et al*. 2023 Deciding what to replicate: a decision model for replication study selection under resource and knowledge constraints. *Psychol. Methods* **28**, 438–451. (doi:10.1037/met0000438)
31. Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, Nosek BA, Johannesson M. 2015 Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl Acad. Sci. USA* **112**, 15343–15347. (doi:10.1073/pnas.1516179112)
32. Arrow KJ *et al*. 2008 The promise of prediction markets. *Science* **320**, 877–878. (doi:10.1126/science.1157679)
33. Tziralis G, Tatsiopoulos I. 2007 Prediction markets: an extended literature review. *J. Predict. Mark.* **1**, 75–91. (doi:10.5750/jpm.v1i1.421)
34. Wolfers J, Zitzewitz E. 2004 Prediction markets. *J. Econ. Perspect*. **18**, 107–126. (doi:10.1257/0895330041371321)
35. Forsell E, Viganola D, Pfeiffer T, Almenberg J, Wilson B, Chen Y, Nosek BA, Johannesson M, Dreber A. 2019 Predicting replication outcomes in the Many Labs 2 study. *J. Econ. Psychol*. **75**, 102117. (doi:10.1016/j.joep.2018.10.009)
36. Gordon M, Viganola D, Dreber A, Johannesson M, Pfeiffer T. 2021 Predicting replicability—analysis of survey and prediction market data from large-scale forecasting projects. *PLoS One* **16**, e0248780. (doi:10.1371/journal.pone.0248780)
37. DellaVigna S, Pope D, Vivalt E. 2019 Predict science to improve science. *Science* **366**, 428–429. (doi:10.1126/science.aaz1704)
38. Landy JF *et al*. 2020 Crowdsourcing hypothesis tests: making transparent how design choices shape research results. *Psychol. Bull*. **146**, 451–479. (doi:10.1037/bul0000220)
39. CORE Team. 2023 Collaborative Open-science and meta REsearch. See http://osf.io/5z4a8.
40. Litman L, Robinson J, Abberbock T. 2017 TurkPrime.com: a versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behav. Res. Methods* **49**, 433–442. (doi:10.3758/s13428-016-0727-z)

41. Palan S, Schitter C. 2018 Prolific.ac—a subject pool for online experiments. *J. Behav. Exp. Financ*. **17**, 22–27. (doi:10.1016/j.jbef.2017.12.004)
42. Benjamin DJ *et al*. 2018 Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10. (doi:10.1038/s41562-017-0189-z)
43. Altmejd A, Dreber A, Forsell E, Huber J, Imai T, Johannesson M, Kirchler M, Nave G, Camerer C. 2019 Predicting the replicability of social science lab experiments. *PLoS One* **14**, e0225826. (doi:10.1371/journal.pone.0225826)
44. Rajtmajer S *et al*. 2022 A synthetic prediction market for estimating confidence in published work. *Proc. AAAI Conf. Artif. Intell.* **36**, 13218–13220. (doi:10.1609/aaai.v36i11.21733)
45. Yang Y, Youyou W, Uzzi B. 2020 Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proc. Natl Acad. Sci. USA* **117**, 10762–10768. (doi:10.1073/pnas.1909046117)
46. Youyou W, Yang Y, Uzzi B. 2023 A discipline-wide investigation of the replicability of psychology papers over the past two decades. *Proc. Natl Acad. Sci. USA* **120**, e2208863120. (doi:10.1073/pnas.2208863120)
47. Chandrashekar SP, Viganola D, Dreber A, Johannesson M, Pfeiffer T, Siegel A, Feldman G. 2025 Using prediction markets to predict replication outcomes of 28 classic articles in social psychology and judgment and decision making. OSF. (doi:10.17605/OSF.IO/2KMH7)
48. Chandrashekar SP, Viganola D, Dreber A, Johannesson M, Pfeiffer T, Siegel A, Feldman G. 2025 Supplementary material from: Using prediction markets and forecasting surveys to predict outcomes of 28 replications in social psychology and judgment and decision making. FigShare. (doi:10.6084/m9.figshare.c.8063063)