Analysis of survey and prediction market data from large-scale replication projects

Authors: Michael Gordon, Thomas Pfeiffer, and Domenico Viganola

Abstract

The credibility of scientific findings is of fundamental importance to enhance future research. One potential approach of collecting information about this credibility is to elicit beliefs about the reproducibility of scientific claims among scientists. Four studies have recently used surveys and prediction markets to estimate beliefs about replication in systematic large scale replication projects, but the sample size in each study has been small. Here we pool data for the four studies (n = 104) to assess the performance of surveys and prediction markets. Both survey beliefs and prediction market beliefs are highly correlated with replication outcomes (correlations > 0.5). Prediction markets predict somewhat better than surveys with lower prediction errors and a higher rate of correct predictions (73% versus 66%). Both prediction markets and surveys suggest that peer scientists are somewhat over-optimistic with average beliefs about 10 percentage units higher than the observed replication rate.

Introduction

While the communication of research findings in scientific publications plays a crucial role in the practice of science, relatively little is known about how reliable and representative the disseminated pieces of information are. Motivated in part by John Ioannidis essay “Why most published findings are false” (Ioannidis, 2005) and by a considerable number of studies that turned out to be false positive (Ioannidis and Doucouliagos 2013, Maniadis et al. 2014), a number of large-scale replication projects were initiated in the behavioral and social sciences (Klein et al. 2014, Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Schweinsberg et al. 2016; Ebersole et al. 2016; Klein et al., 2018; Cova et al. 2018). These projects aimed to evaluate a large sample of studies from specific research fields through direct replication, which involves the collection and analysis of novel experimental data with methodology as similar as possible to the original study. This means that original studies were redone with similar materials in larger samples and new participants. Replication rates ranged from 39% to 62%, with the main replication indicator defined as an effect in the same direction as in the original study and statistically significant (typically $p < 0.05$ in a two-sided test, though the Many Labs projects had a different cutoff).

Four of these systematic replication projects were accompanied by prediction markets and prediction surveys aimed at forecasting the replication outcomes before the replications were conducted (Camerer et al., 2016, 2018; Dreber et al., 2015; Forsell et al., 2018). The purpose of these prediction market studies was to investigate whether there is information within the research communities about which studies are likely to replicate, and which ones are not; and whether this information can be elicited with prediction markets or surveys. In this paper, we pool the data from these four studies that taken together elicited peer beliefs about the replication outcomes of 104 published studies, mainly in the fields of experimental psychology and experimental economics. Overall, 51 studies out of 104 replicated.[1] By pooling the data across these four studies we get substantially more statistical power to test the performance of the prediction markets and surveys than in the individual studies based on relatively few observations. We find that both the prediction market beliefs and the survey beliefs are highly correlated with a successful replication, with correlations over 0.5. In a typical prediction markets platform, the participants can buy and sell contracts that guarantee a fixed amount if a specific event occurs and zero otherwise. The final prices of these contracts are endogenously determined by the trading activity within the markets, they are bounded between 0 and 1, and they can be interpreted as the aggregate beliefs that the events (i.e., the objects of the trades) will occur. In our framework, these aggregate beliefs can be used as a proxy of the predicted probabilities that the selected studies will replicate. Interpreting a predicted replication probability over 50% as predicting a successful replication and a predicted replication probability below 50% as predicting a failed replication, we find that the prediction markets correctly predict 76 of the 104 studies (73%), and that the survey correctly predicted 68 out of 103 studies (66%).[2] In line with this moderately higher prediction accuracy of the markets, the prediction markets were also associated with somewhat lower prediction errors than the survey. This may be due to the larger spread in predictions obtained with the prediction markets. As probabilities are bounded between 1 and 0, the noise in individual-level survey predictions bias the aggregate forecasts towards a probability of 0.5 reducing the spread of the final predictions (Atanasov et al., 2017). We furthermore find that both the prediction markets and the survey

[1] Refer to table 2 for a detailed report of project-specific studies are replication rates.

[2] Data for the survey is missing for one of the 104 replication studies; refer to the 'RPP' paragraph in the description section for further details.

overestimate the average replication probability by about ten percentage units, suggesting that peer scientists are somewhat over-optimistic about the credibility of the original studies.

Below we describe the common methodology used in the four prediction projects and we summarize the main findings in the individual projects. We thereafter present the results based on the pooled data, and end with some concluding remarks.

Description of the 4 projects

This paper analyzes the resulting dataset from pooling the data of four published projects comparing prediction markets and survey. These four projects ordered according to when they were published are: Dreber et al 2015 (also referred as RPP in the rest of the paper), Camerer et al. 2016 (EERP), Forsell et al. 2018 (ML2), and Camerer et al. (SSRP).[3] For the sake of clarity, in the rest of the paper we will refer to the above mentioned papers as 'projects', while the terms 'studies' or 'claims' will refer to the papers selected for replication in each project. Therefore, this paper analyzes the dataset pooling the observations from 4 *projects* about 104 *studies*.

Common Methodology across the 4 projects

The prediction market projects (RPP, EERP, ML2, and SSRP) were focused on the outcome of replications of scientific claims published in academic journals. These four projects answer the same question in different fields of the social sciences: can we use prediction markets to detect which published studies are more likely to replicate? Typically, one key finding of a publication was selected to be replicated with a methodology as close as possible to the original paper, with replication power typically exceeding the statistical power of the original finding. All the projects shared a similar structure, and the participants' forecasts were elicited in a similar way across the 4 projects.[4] Before the replication outcomes became public information, peer researchers first participated in a survey eliciting beliefs about the replication probability and thereafter participated in prediction markets. The prediction markets were designed to predict the probability of successful replication. Within a prediction market, participants were endowed with tokens that could be used to buy and sell contracts that paid one token if a finding

[3] In RPP and ML2 the authors relied on external projects for the selection of the studies and for conducting the replications (Open Science Collaboration 2015 and Klein et al. 2018, respectively). In the EERP and SSRP all the steps of the analysis were managed by the authors.

[4] Refer to table A1 in the appendix for further details about the project-specific recruitment process of the forecasters.

was replicated, and 0 tokens if it was not replicated (and at the end of the study tokens were converted to US dollars at an exchange rate of 1 or 0.5 in the four different studies). A successful replication was defined as a significant effect in the same direction as in the original study. The emerging price for such a contract can be interpreted as a collective forecast of the probability of a study replicating, albeit with some caveats (Manski 2006).[5] An automated market maker implementing a logarithmic market scoring rule was used to determine prices (Hanson, 2003). The prediction markets lasted for 2 weeks in the RPP, the ML2, and the SSRP, and for 10 days in the EERP. Some participants who filled out the survey did not participate in the prediction markets, but all data below for the survey are based on participants who actively participated in the markets (i.e. a participant had to trade in at least one market to be included in the survey data); this is so that both the survey and prediction market data is based on the same participants.[6] However, as the survey data are not available for one study of the RPP project, in this paper we analyze data for 104 prediction markets and 103 prediction surveys.

For all the projects, the authors of the original studies were contacted and asked to provide feedback on the replication designs before starting the data collection for the replications. For the RPP, the EERP, and the SSRP a replication is deemed successful if it finds a 'significant effect size at 5% in the same direction of the original study' (Open Science Collaboration 2015, Cumming, 2008); for the ML2 a replication is deemed successful if it finds 'a significant effect size in the same direction of the original study and a p-value smaller than 0.001'. The latter definition of a successful replication is more stringent because the power of the replications in the ML2 project is higher with the multiple laboratories data collections. The average (mean) survey response was calculated for each study, and in the following analyses, we refer to it as the 'survey belief'. Once the market was closed and all replications had been performed, market forecasts and survey belief were compared against the replication results to evaluate the predictive power of the market and survey. Further characterizations of each project are summarized in table 1:

Table 1: main features of individual projects

	RPP	EERP	ML2	SSRP

[5] Refer to Dreber et al. 2015 and the references therein for a detailed explanation of the functioning of prediction markets.

[6] However, the four individual studies present results also for all the survey participants. These results are very similar to those obtained focusing only on the participants that were also active in the markets, suggesting that the participants did not self-select into the markets based on their survey performances.

Field of study	Experimental Psychology	Experimental Economics	Experimental Psychology	Experimental Social Science
Source Journals	JPSP, PS, JEP (2008)	AER, QJE (2011-2014)	Several psychology outlets, including JEP, JPSP, PS[7] (1977-2014)	Science, Nature (2010-2015)
Claim selected	In the case of several studies in one paper, typically the last study of each paper was selected for replication	4 criteria in descending order: select the most central result in the paper (among the between-subject treatment comparisons) based on to what extent the results were emphasized in the published versions; if more than one equally central result, the result (if any) related to efficiency was picked, as efficiency is central to economics; if several results still remained and they were from different separate experiments the last experiment (in line with RPP) was picked; in case several results still remained one of those results was randomly selected for the replication	Selected by the authors of the Many Labs 2 project, with the aim of assuring diversity and plurality of claims	3 criteria in descending order: select the first study reporting a significant effect; select the statistically significant result identified as the most important result; random selection in case of more than one equally central result
Rule to define the replication sample size based on power calculations	Replication teams required to propose a study design that would achieve at least 80% power to detect the original effect size and were encouraged to obtain higher power if feasible (average proposed power: 92%)	At least 90% power to detect the original effect size at 5% significance level	Sample size in each study was expected to be over 4500 participants, implying that the power of the replications to detect the original effect size was expected to above 99%. In the market intro, the participants were told that 'all studies have a power close to 100%'	Stage1: 90% power to detect 75% of the original effect size at 5% significance level Stage 2: 90% power to detect 50% of the original effect size at 5% significance level. This procedure lead to sample sizes on average about 5 times bigger than the original studies

The RPP and EERP had a similar power of about 90% to find the original effect size. But as effect sizes in original studies of true positive findings may be inflated, these studies may in fact be underpowered on average. Therefore the power was increased substantially in the SSRP

[7] Refer to Klein et al. 2018 for a detailed description of the selection procedure for the studies to be included in the Many Labs 2 project.

to have 90% power to detect 75% of the original effect sizes in the first stage and the same power to detect 50% of the original effect size in the second stage. In the ML2 the power is even higher due to the very large sample sizes across multiple sites. The information about power of the original studies and of the replications was disclosed to the forecasters participating in predictions elicitation phase (i.e., prediction markets and surveys). In addition, the most relevant information (including the power of the replications) was embedded in the survey and in the market questions, the links to the original publications were provided, and, when available, the forecasters were also provided with the pre-replication versions of the replication reports detailing the design and planned analyses of each replication. The rest of this section provides a summary of the projects; the interested reader should refer to the original publications for further details.

RPP - Using prediction markets to estimate the reproducibility of scientific research

Dreber et al. 2015 rely on a large scale replication project (Open Science Collaboration 2015; replication rate: 39%) and identify a set of studies published in the Journal of Personality and Social Psychology, Psychological Science, and Journal of Experimental Psychology suitable for eliciting beliefs on the likelihood of successful replication. They run and evaluate 41[8] prediction markets and 40 surveys in two separate batches in November 2012 and in October 2014 to study whether researchers beliefs carry useful information about probability of successful replication: they conclude that the prediction markets (surveys) correctly predicted the outcome of the replications 71% (58%) of the times. The Spearman correlation between the prediction markets final prices and the replication outcome was 0.423 (p = 0.006), while the one between the average survey beliefs and the replication outcome was 0.244 (p = 0.130).

EERP - Evaluating replicability of laboratory experiments in economics

The focus of Camerer et al. 2016 is on experimental economics: 18 studies published in two of the top-5 economic journals (American Economic Review and Quarterly Journal of Economics) were replicated, with the share of successful replications ranging between 61% and 83% and a mean relative effect size of the replications to the original studies being 66%. All the studies were also the objects of prediction markets and surveys with the aim of eliciting beliefs on which studies would replicate: both the markets and the survey correctly categorized 11 studies out of 18 (61%). In the EERP, the Spearman correlation between the prediction

[8] Originally 44 studies were selected but only 41 replications were completed, thus only 41 markets cleared. The survey data about one of the studies in the RPP is not available as the framing of the question in the survey did not match the claim that were selected for replication for that study, thus the answers could not be used for the analysis. This issue was fixed for the subsequent prediction markets sessions, which are therefore included.

markets final prices and the replication outcome was 0.297 (p = 0.232), while the one between the average survey beliefs and the replication outcome was 0.516 (p = 0.028).

ML2 - Predicting replication outcomes in the Many Labs 2 study

Forsell et al. 2018 lean on a large replication project, the Many Labs 2, lead by the Open Science Collaboration. One of the aims of the Many Labs 2 study was to guarantee high-quality standards for the replications of classic and contemporary findings in psychology, achieving large sample sizes across different cultures and labs, and requiring replication protocols to be peer-reviewed in advance. The realized replication rate for the ML2 project is 45.8% (11 successful replications out of 24 studies analyzed). In addition to the prediction markets and the prediction surveys on the binary outcome 'study *x* will replicate or not', Forsell et al. 2018 also ask participants to submit their beliefs about the realized effect size in the replications.[9] They focus on 24[10] studies: the prediction markets (surveys) correctly predicted 75% (67%) of the replication outcomes. In the ML2 project, the Spearman correlation between the prediction market final prices (surveys) and the replication outcome was 0.755 (0.731), with p < 0.001 in both cases.

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

Camerer et al. 2018 set up a replication project targeting 21 social science experimental studies published in two general science outlets: Science and Nature. In particular, the SSRP is designed to address the issue of inflated effect sizes in original studies. In line with previous projects, also Camerer et al. 2018 run prediction markets and prediction surveys to forecasts whether the selected studies will replicate. However, differently from the three previous projects, the SSRP design for conducting replications is structured in two stages: first, it considers 90% power to detect 75% of the original effect size; if the replication fails, stage 2 starts and the data collection keeps running until reaching 90% power to detect 50% of the original effect size (pooling data from stage 1 and stage 2 collection phases). Based on all the data collected, 62% of the 21 studies successfully replicated. The prediction markets follow a similar structure of the data collection: participants were randomized in two groups: in treatment 1 beliefs about replicability in stage 1 were elicited; in treatment 2 beliefs about

[9] Results about the effect sizes prediction markets/surveys fall outside the scope of this paper, thus are not discussed. Refer to Forsell et al. 2018 for the detailed analysis.

[10] Although the Many Labs 2 replicated 28 studies, the prediction markets and the surveys were analyzed for a subset of 24 studies as four of the original studies focused on cultural differences in effect sizes across different samples. Note that when including all the 28 studies, the replication rate of the Many Labs 2 project (Klein et al. 2018) increases to 50% (14/28). Refer to Appendix A of Forsell et al. 2018 for further details.

replicability in *both* stage 1 and stage 2 were elicited. In this paper, we report the results about treatment 2 only, as the replication results after Stage 2 is most informative about the replication outcome. In SSRP, both the prediction markets and the survey correctly predicted 86% of the replication outcomes, however the Spearman correlation between the prediction markets final prices and the replication outcomes (0.842, p < 0.001), was higher than the Spearman correlation between the average survey beliefs (0.761, p < 001).

Analysis of the pooled data

Descriptive statistics

The first four columns of table 2 report the relevant statistics about the replication rates and the accuracy of prediction markets and prediction surveys disaggregated by project. The last column shows the statistics obtained when pooling the data from the 4 projects; more detailed information about the markets can be found in table A1 of the appendix. While this section is mainly descriptive, the formal statistical tests employed to investigate the performances of the prediction markets compared to the performances of the prediction surveys are the object of the next section.

For the prediction markets, we interpret the final price of each claim as the elicited probability that the claim would replicate, in line with the 4 original projects aforementioned. In particular, if the final price exceeds 0.50, we interpret that the market predicts a successful replication; if the final price is lower than 0.50, we interpret that the market predicts a failed replication. The same rules apply for surveys: we compute the average beliefs for each study and then interpret that the survey predicts a successful replication if the average beliefs exceed 0.50 and a failed replication otherwise.

For each study in each project, we compute the one-dimension euclidean distance between the forecasted outcomes and the realized outcomes and refer to it as the absolute prediction error (APE). The absolute prediction error associated to the prediction markets is computed as: $APE_{ip}^{pm} = |f_{ip} - O_{ip}|$ where f_{ip} is the final price for study *i* in project p and $O_{ip}$is the realized outcome in terms of successful replication ($O_{ip} = 1$) or failed replication ($O_{ip} = 0$) for study *i* in project p. Accordingly, the absolute prediction error associated to the prediction surveys is computed as $APE_{ip}^{su} = |\underline{b}_{ip} - O_{ip}|$ where $\underline{b}_{ip}$ is the average belief elicited through the survey for study i in project p.

Table 2: market and survey results - subdivided by projects

	RPP	**EERP**	**ML2**	**SSRP**	**Pooled data**
N. studies	41[11]	18	24	21	104
Successful replications	16	11	11	13	51
Replication share	39.0%	61.1%	45.8%	61.9%	49.0%
Correct PM (%)	29 (71%)	11 (61%)	18 (75%)	18 (86%)	76 (73%)
Mean beliefs PM	0.560	0.751	0.644	0.634	0.627
Range beliefs PM	0.132 - 0.879	0.588 - 0.937	0.271 - 0.923	0.231 - 0.955	0.132 - 0.955
Mean APE PM	0.427	0.414	0.354	0.303	0.383
Correct Survey (%)	23 out of 40 (58%)	11 (61%)	16 (67%)	18 (86%)	68 out of 103 (66%)
Mean beliefs survey	0.546	0.711	0.647	0.605	0.610
Range survey	0.339 - 0.740	0.542 - 0.863	0.327 - 0.887	0.278 - 0.812	0.278 - 0.887
Mean APE Survey	0.485	0.409	0.394	0.348	0.423
Spearman Correlation - PM and Survey beliefs	0.736	0.792	0.947	0.845	0.837

From table 2 it emerges that the share of successful replications ranges from 39% to 62%, with an overall rate slightly below 50%. In terms of shares of correct predictions, the surveys never outperform the markets: in two cases (EERP and SSRP) they correctly categorize the same number of studies in the replicates/non-replicates dichotomy; in the other two projects the markets do better (71% vs 58% in the RPP; 75% vs 67% in the ML2). Overall, the prediction markets are correct 73% of the times (76 out of 104 studies), while the prediction surveys are correct 66% of the times (68 out of 103 studies). In terms of mean absolute prediction error, the markets are more accurate when compared to the surveys for three of the four projects, with the exception of the EERP where the absolute prediction error is slightly lower for the survey.

[11] 41 studies analyzed for the prediction markets; 40 studies analyzed for the prediction surveys. Note that the study excluded in the surveys did replicate. Refer to Table A2 in the appendix for a focus on the 103 studies for which both prediction markets data and prediction surveys data are available.

The Spearman correlation is high between markets and survey beliefs in all the four projects ranging between 0.736 and 0.947.

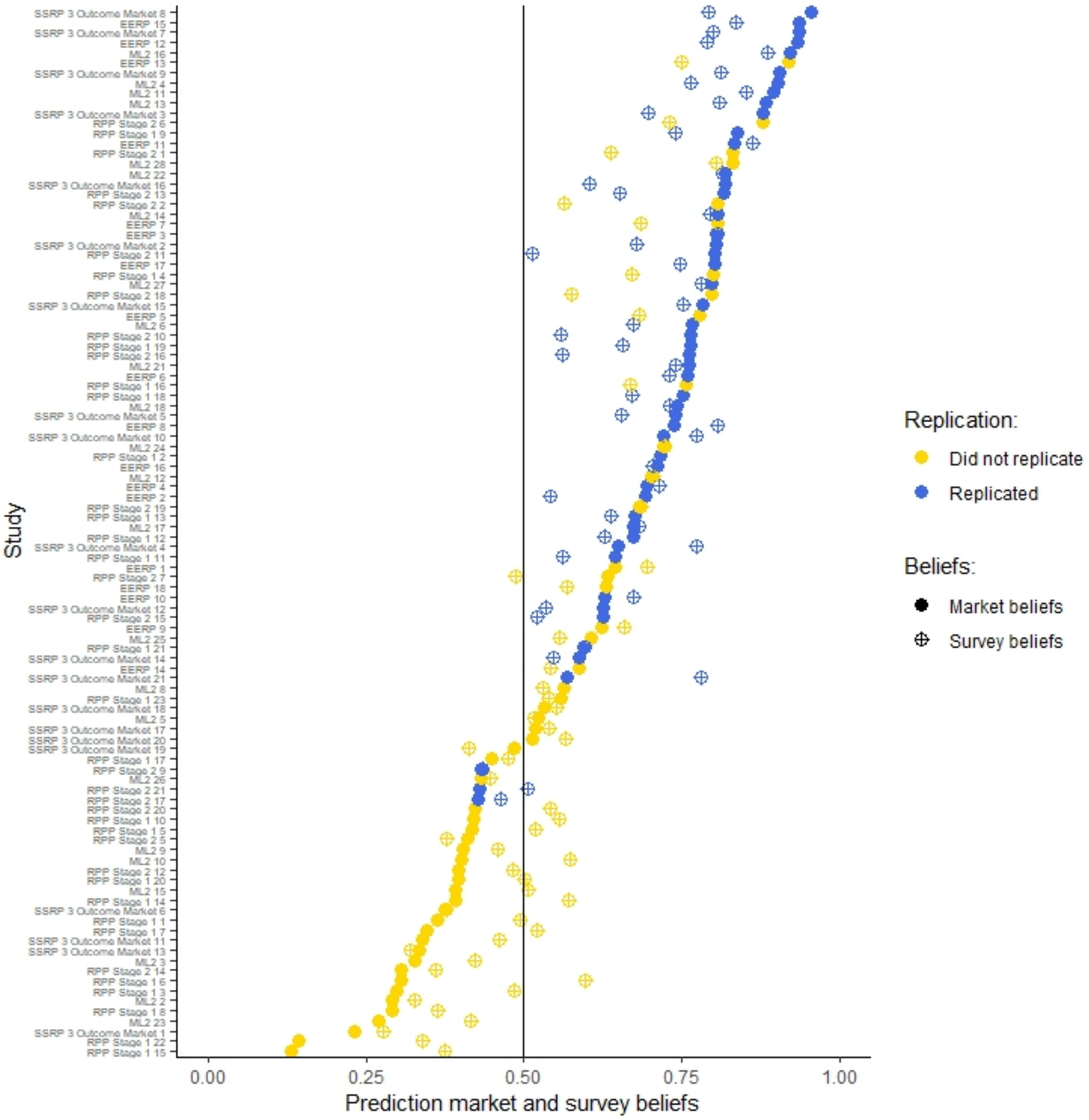
The mean beliefs are similar for the markets and the survey in all four projects, although slightly higher for the former. There is a tendency for more dispersion in the market beliefs than in the survey beliefs: the ranges between the lowest and the highest final price are wider in all four projects than the ranges between the corresponding survey beliefs. Wider ranges are consistent with the idea that prediction markets tend to be more polarized towards the extremes of the likelihoods, while surveys tend to be flattened around the mean, suggesting that the markets have higher discriminatory power.

Statistical Analysis

In this section, we report and comment on the outcomes of the statistical analyses performed to compare the prediction markets results and the survey results in a systematic way. For each hypothesis, we specify two versions of the same test: a parametric version and its non-parametric equivalent. This approach is justified by observing that all the tests are performed on more than 100 observations, thus we consider the standard assumptions of parametric tests to be fulfilled. However, in order to guarantee that our results are comparable with those reported in Dreber et al 2015, Camerer et al. 2016 and 2018, and Forsell et al. 2018, we also report the non-parametric equivalents. For all the results reported below, the tests should be interpreted as two-tailed tests and a p-value < 0.005 should be interpreted as “statistically significant” while a p-value < 0.05 as “suggestive” evidence, in line with the recommendation of Benjamin et al. (2018).

As reported in table 2, in the pooled data for the four projects the prediction markets correctly identify which studies will replicate 76 times out of 104 (73.1%), while the surveys have a lower rate of correct predictions: 68/103 (66.0%). In both cases, the aggregate beliefs are deemed to be in favor of successful replication if they exceed the threshold of 0.50 (solid vertical line in figure 1).

Figure 1: Prediction market and survey beliefs for the successful replication probability. The figure shows the beliefs elicited through prediction markets - filled dots - and through surveys - hollow dots with a cross - for each of the studies in the pooled dataset. The replication studies on the y-axis are ranked in terms of final prices of the prediction markets, with studies less likely to replicate at the bottom and the studies more likely to replicate at the top. Both the prediction markets beliefs and the survey beliefs are highly correlated with successful replications (Spearman correlations = 0.567, p < 0.001, n = 104 for the



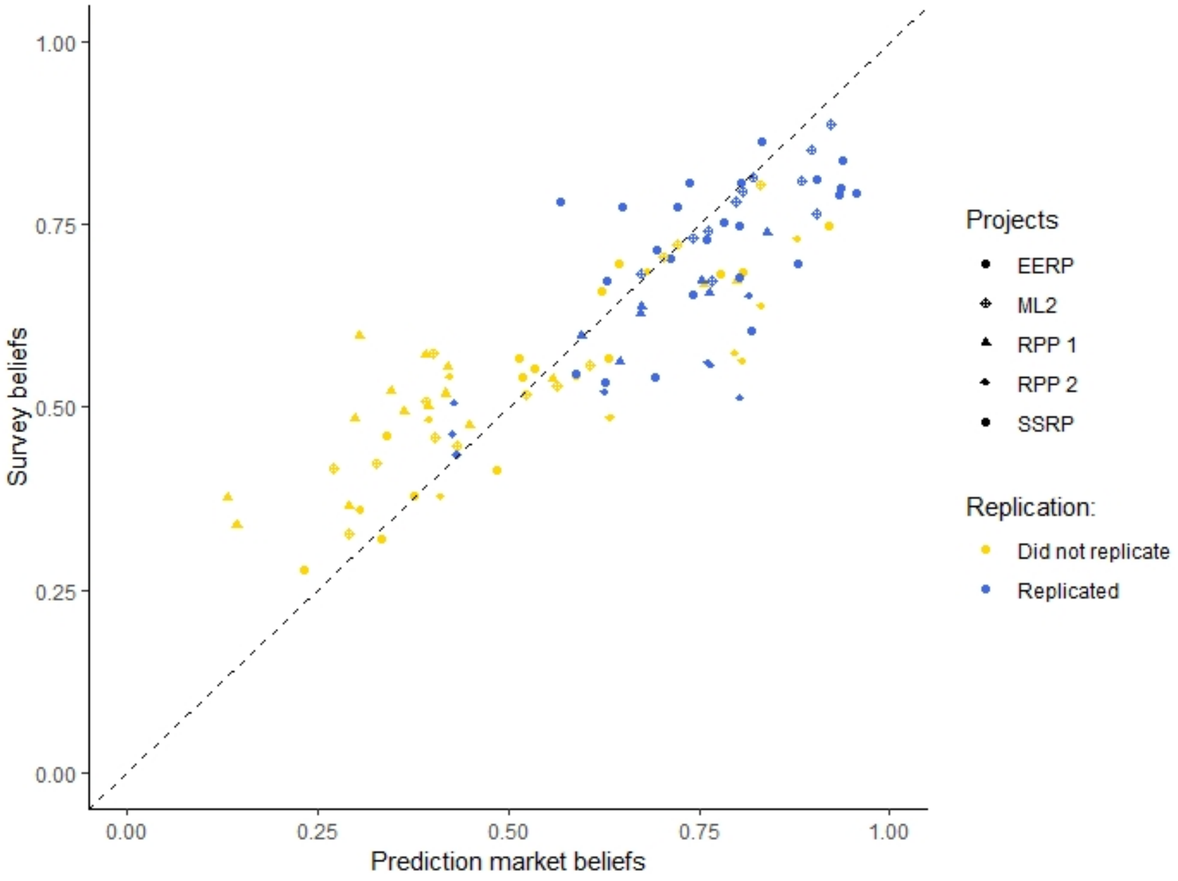
prediction markets and 0.557, p < 0.001, n = 103 for the survey). Note that survey data are missing for study 2 in the RPP stage 1 project.

As figure 1 shows, studies that do not replicate (yellow) tend to lie on the left of the 0.50 threshold, while studies that do replicate (blue) are more concentrated on the right of the threshold. In particular, out of the 31 studies that are predicted by the market not to replicate, only 3 eventually replicate, thus for these studies the market is correct more than 90% of the times. On the other hand, out of the 73 studies that are predicted to successfully replicate, 25 did not replicate, with a correct prediction rate of 65.8%. The shares of correct forecasts branched by whether the predictions suggest a failed replication or a successful replication are quite similar for the surveys: 90.9% and 59.3% respectively (out of the 22 studies that are predicted not to replicate by the survey, only 2 eventually replicate; out of the 81 studies that are predicted to successfully replicate, 33 do not replicate). Both the markets and the surveys are more accurate when concluding that a study will not replicate rather than when concluding that a study will replicate. This may at least partially be due to the limited power of the replications in RPP and EERP, as some of the failed replications may be false negatives (the power of the replications puts an upper bound on the correct prediction rate for studies predicted to replicate).

For both the prediction markets and the survey, we test by means of a one-sample binomial test if the fraction of correct predictions is statistically different from the 50% threshold, which is the success rate we would expect with a flat prior and with equal probabilities of successful and unsuccessful replication, i.e., the success rate one would get by pure randomness tossing a coin to determine if a study will replicate or not. We find that the rates of correct predictions of both the prediction markets and of the survey are statistically different from the 50% threshold (one-sample binomial test: $p < 0.001$, $n = 104$ for the prediction markets and $p = 0.001$, $n = 103$ for the prediction survey), suggesting that aggregating beliefs generate useful information to detect which studies are more likely to replicate.

Survey beliefs and prediction markets beliefs are highly correlated (Spearman correlation test = 0.837, $p < 0.001$; Pearson correlation test = 0.853, $p < 0.001$ with $n = 103$ for both tests); as it emerges distinctly from figure 2. Moreover, the fact that market and survey beliefs are highly correlated is not driven by the studies of a particular project, rather it is a feature observable for all the studies and across all the projects.

Figure 2: Correlation between market and survey beliefs across projects, n = 103.

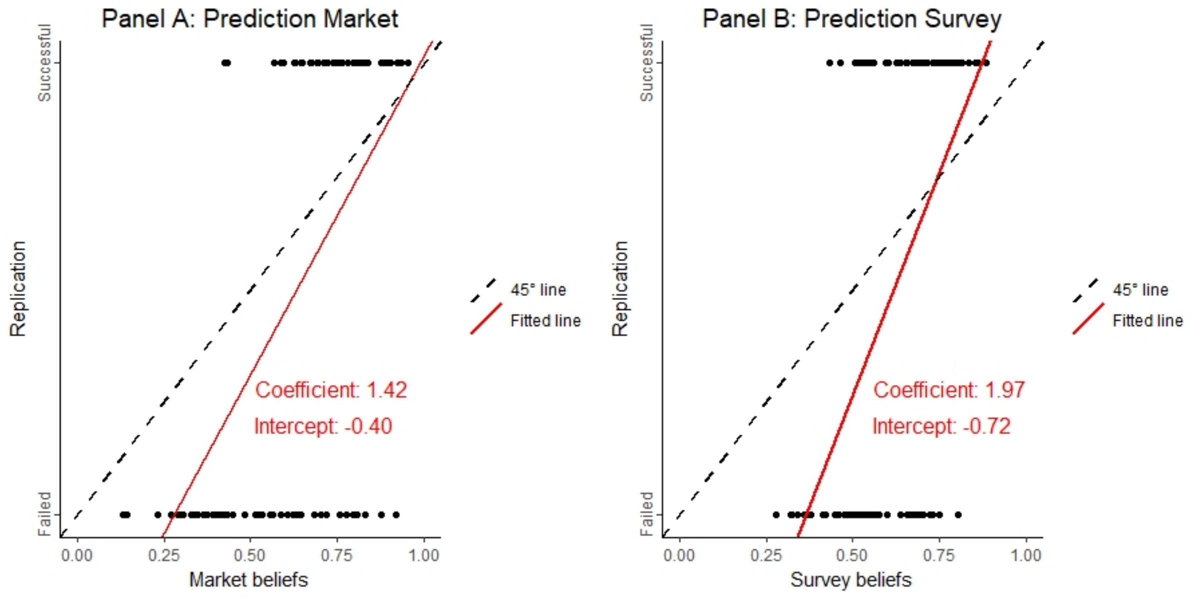


When identifying correct predictions with final prices being above 0.50 (for the prediction markets) and average beliefs being above 0.50 (for the survey) for successful replications and below the same thresholds for failed replications, the prediction markets are correct in 8 additional cases with respect to the surveys (76 out of 104 correct forecasts for the markets, 68 out of 103 for the survey). But is this difference statistically significant? Do the prediction markets outperform the survey when looking at the shares of correct predictions? We answer this question both non-parametrically (via Wilcoxon signed-ranks test between the correct predictions of the prediction markets and the correct predictions of the survey) and parametrically (via paired t-test between the same vectors). Both tests find suggestive evidence, but not statistically significant evidence, that prediction markets perform better than surveys (mean of the differences = 0.068, Wilcoxon signed-rank test using Pratt's method to account for zero values, $p = 0.039$; paired t-test with df = 102, $p = 0.034$; n = 103 in both cases).

Next, we investigate whether the prediction market and the survey forecasts are well calibrated. Given that the replication rate obtained pooling all the studies is 49%, a well-calibrated forecasting method should predict that half of the studies replicate and half do not. However, both the average prediction markets beliefs (0.627) and the average survey beliefs (0.610) are higher than the realized replication rate. Thus in order to attest whether the two beliefs elicitation methods are well calibrated, we test if the final prices and the average survey beliefs over-estimate the actual replication rates. Both the non-parametric test and the parametric test find evidence in favor of overestimation for the prediction markets (Wilcoxon signed-ranks : p < 0.001, paired t-test : p = 0.001, n = 103). For the survey, while the non-parametric test finds statistical evidence in favor of over-estimation, the parametric test finds only suggestive evidence (Wilcoxon signed-ranks: p < 0.001, paired t-test p = 0.005).

Although overestimating the true replication rates, both the prediction market prices (Spearman correlation = 0.567 , p < 0.001; Pearson correlation: 0.582, p < 0.001) and the survey beliefs (Spearman correlation = 0.557 , p < 0.001; Pearson correlation: 0.564, p < 0.001) are highly correlated with the replication outcomes, suggesting that there is scope for adjusting the estimated final prices and to achieve higher calibration.

Figure 3: linear probability model for the prediction markets (A) and the survey (B)



The correlation between the realized replication rates and the prediction markets/survey forecast can be further assessed by regressing the dummy variable identifying successful replication on the final prices from the markets (panel A in figure 3) and on the average beliefs elicited through the surveys (panel B in figure 3). The full regression tables are summarized in table 3. For the prediction markets, the coefficient of the independent variable is β= 1.415, t(102) = 7.23, CI [1.027, 1.804], p < 0.001; for the survey, the corresponding coefficient takes value β = 1.973, t(101) = 6.86, CI [1.400, 2.544], p < 0.001. Ideally, if the market prices and the survey averages can be interpreted as probabilities of replications, one would expect the coefficient of the independent variable to be $\beta \approx 1$, and the intercept to be close to zero. For the prediction markets, while the slope coefficient is statistically different from zero, there is only suggestive evidence that it is also different from one (p-value = 0.036). The intercept = -0.379 however is statistically different from zero (p = 0.003). On the other hand, the slope coefficient relative to the survey beliefs (column 3) is statistically different both from 0 (p < 0.001) and from 1 (p < 0.001), and the intercept = -0.719 is statistically different from 0 (p < 0.001). These results suggest that there is scope for improving the calibration of both the prediction markets and the prediction surveys.[12]

Table 3. Linear and logistic regressions for prediction markets and survey outcomes.

	Dependent Variable: realized outcome	
	Linear model	
	Markets	Surveys
	(1)	(2)
Intercept	-0.397**	-0.719**
	(0.129)	(0.180)
Beliefs	1.415**	1.973**
	(0.196)	(0.288)
N	104	103
$Adj - R^2$	0.332	0.311

Notes: '**' p-value < 0.005; '*' p-value < 0.05. Standard errors in parenthesis.

Analysis of error rates

[12] In order to accommodate the fact that the dependent variable is a dummy, we also fit a logistic model. In a logistic model there is a linear relationship between the log of the odds ratio and the estimated regression equation. In the logistic model we therefore transform the independent variable (the market belief or the survey belief) to the logarithm of the odds ratio. With this specification the coefficient of the independent variable should equal 1 if the market beliefs (survey beliefs) can be interpreted as probabilities. The estimated coefficient in the regression using the log of the odds ratio of the market beliefs is 1.545 (se = 0.321) which is not significantly different from 1 (p=0.084); the estimated coefficient for the survey beliefs is 2.412 (se = 0.489) which is significantly different from 1 (p = 0.004).

Another standard way to assess forecasting accuracy is by determining the absolute prediction errors of the forecasts. The average prediction errors of the prediction markets (APE^{pm}, mean = 0.383, median = 0.343, range = [0.045; 0.920], n = 104) is lower than the average prediction error associated to the surveys (APE^{su}, mean = 0.423, median = 0.438, range = [0.113; 0.804], n = 103). In particular, in 70 cases out of 103, the absolute prediction error associated to the prediction markets is lower if compared to the absolute prediction error associated with the survey. A non-parametric test between APE^{pm} and APE^{su} rejects the null hypothesis of the difference of the means being equal to zero (difference of means = -0.039, Wilcoxon signed-ranks $p < 0.001$, n = 103). This result is aligned to the parametric paired t-test between the same variables ($p < 0.001$). Both these results are confirmed by the distributions of the APE^{pm} and APE^{su}shown in figure 4.

The accuracy of the forecasts can also be measured in terms of the Brier score (Brier 1950), a proper scoring rule ranging between 0 and 1. Higher levels of the Brier score are associated to worst forecasts, while the value 0 is obtained when the forecast matches exactly the outcome of the probabilistic event. In particular, the Brier score is computed as the mean squared difference between the predicted probabilities assigned to a probabilistic event and the actual outcome of that event: as in this paper we are dealing with binary events, the Brier score for each study in the prediction markets is computed as $B_{ip}^{pm} = (f_{ip} - O_{ip})^{2}$, while for the survey it is computed as $B_{ip}^{su} = (\underline{b}_{ip} - O_{ip})$.^{2}

The difference between the accuracy rates of the markets and of the survey is less remarkable when using the Brier score rather than the absolute prediction errors measured using APE^{pm} and $APE^{su}_{.}$ The reason being that the Brier score penalizes more incorrect and extreme forecasts, and as shown before, markets tend to produce more polarized forecasts. Analytically, the average Brier score across all the prediction markets is 0.191, while the average Brier score across all the surveys is 0.205. The difference between the two means (0.013) is statistically different from zero when tested non-parametrically (Wilcoxon signed rank test: p = 0.0045) but it is not statistically different from 0 when tested with a parametric test (paired t-test: p = 0.205). The distribution of the Brier score for the market and survey beliefs are shown in figure 5.

Figure 4. Distributions of absolute prediction errors: Survey VS Market beliefs

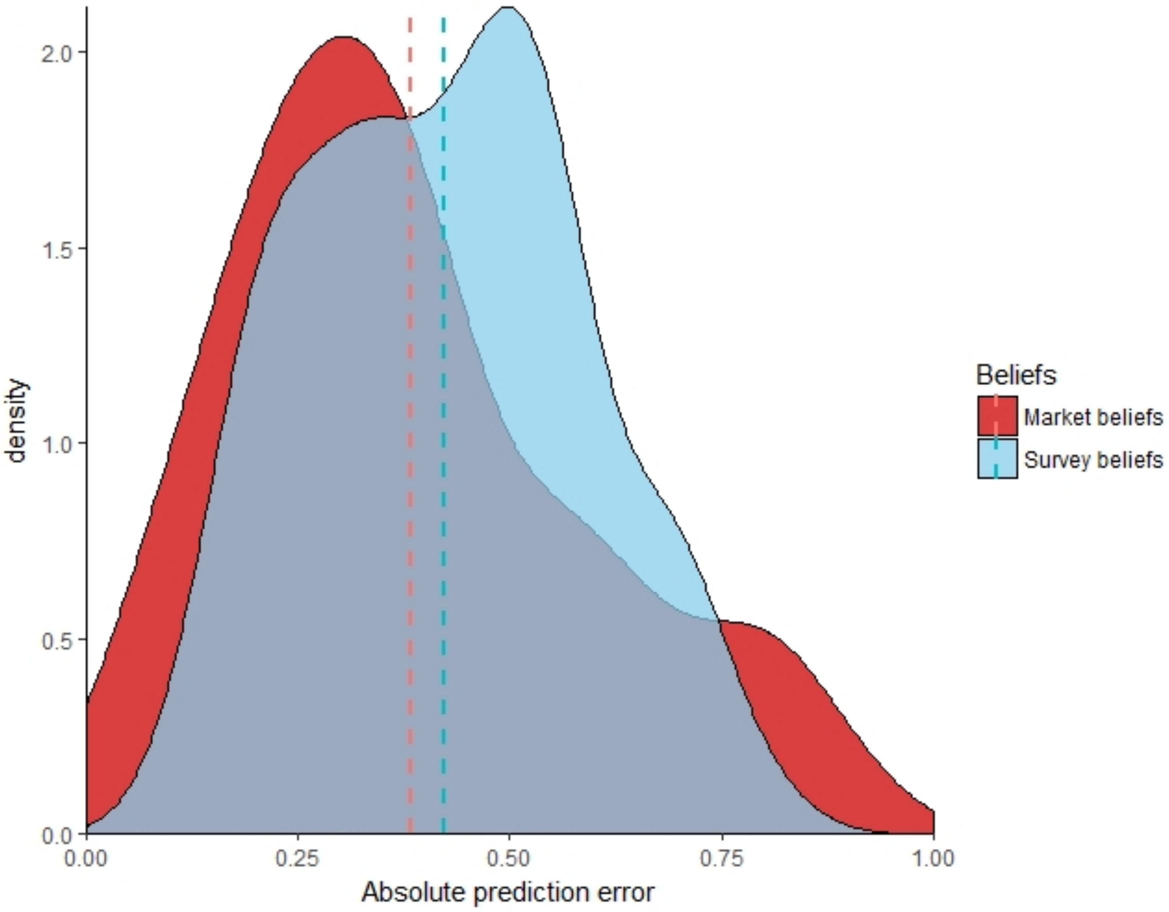
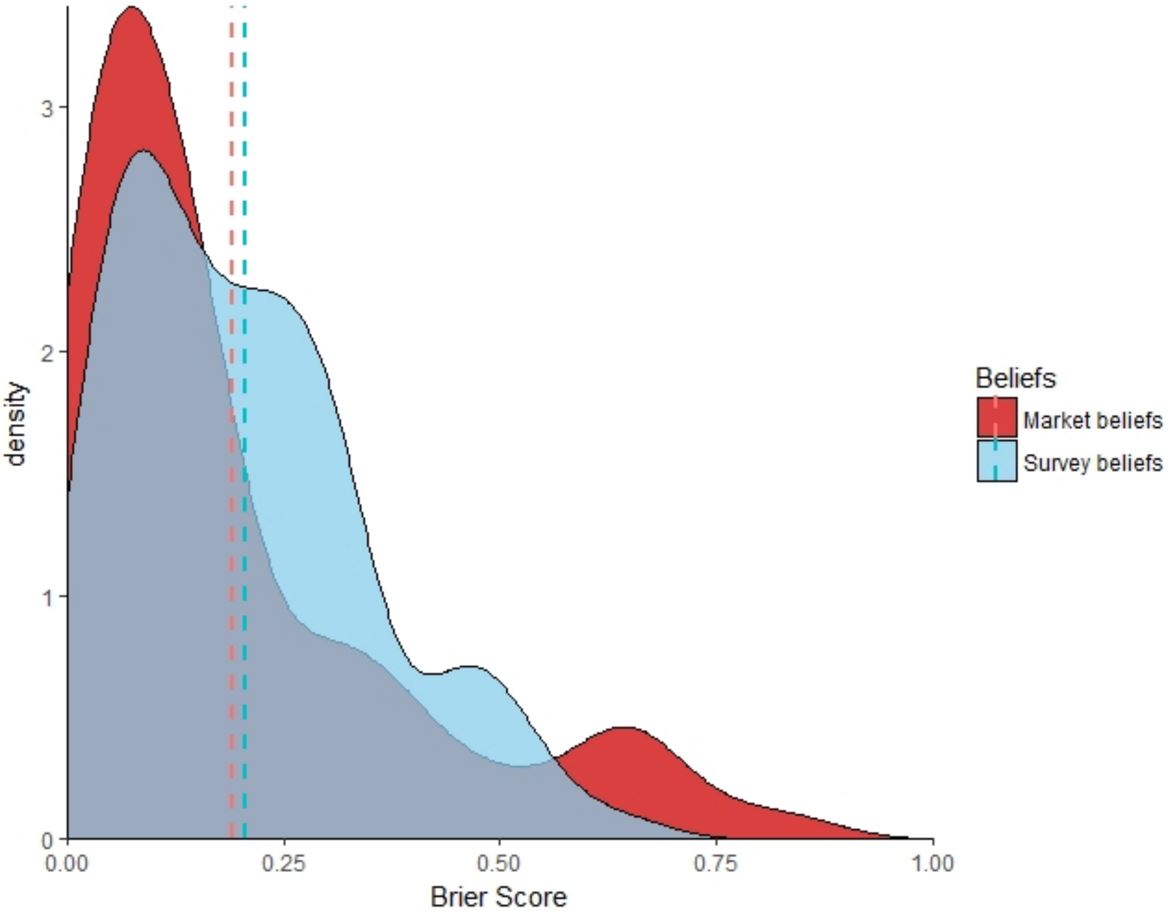


Figure 5. Distributions of Brier scores at study level: Survey VS Market beliefs



Conclusions

In this paper, we provide an analytical investigation of the forecasting performances obtained by two different procedures to elicit beliefs about replication of scientific studies: prediction markets and prediction survey. We pooled the forecasting data using these two methods from four published papers in which forecasters, mainly researchers and scholars in the social sciences, had to estimate the likelihood that a tested hypothesis taken from a paper published in scientific journals would replicate. We find that the prediction markets correctly identify which studies successfully replicate and which do not 73% of the times (76/104), while the prediction surveys are correct 66% of the times (68/103). Both the prediction market estimates and the prediction surveys estimates are highly correlated with the replication outcomes of the studies selected for replication (Pearson correlation = 0.582 and = 0.564, respectively), suggesting that to some extent, studies that replicate are systematically different and identifiable from studies that do not successfully replicate. However, both the forecasts elicitation methods tend to overestimate the realized replication rates, and beliefs about replication are on average about ten percentage units larger than the observed replication rate. The results suggest that peer beliefs can be elicited to obtain important information about reproducibility, but the systematic overestimation of the replication probability also imply that there is room for calibrating the elicited beliefs to further improve predictions.

In terms of comparing which elicitation method performs better in the task of aggregating beliefs and providing more accurate forecasts, our results suggest that the markets perform somewhat better than the survey. There is suggestive evidence for a higher rate of correct predictions for market beliefs, and the absolute prediction error is significantly lower for the markets. The comparison is less clear-cut using the Brier score that squares the prediction errors; the Brier score is still lower for prediction market beliefs than for survey beliefs but the difference is smaller and only significant for the non-parametric test and not for the paired t-test. The difference in results for the absolute prediction error and the Brier score is shown in Figures 4 and 5: the prediction markets have a tail of high prediction errors, that are more heavily weighted in the Brier score. These are due to some studies where the markets predicted very high likelihoods of replication, but these studies failed to replicate. Note however that it is possible that some of these studies are false negative replication results, due to insufficient replication power.

As suggested in Plott and Sunder 1988, prediction markets are potentially more accurate than surveys as within a prediction markets framework, the information embedded in the current prices is shared with all the active traders and for the whole duration of the markets, allowing traders to update their beliefs in light of other participants' beliefs. Moreover, traders can decide how to allocate their endowments across different claims, thus they have the freedom to bet only when they are particularly confident about a specific claim. With this regard, a crucial difference between prediction markets and prediction surveys is that in the first case a 'wrong' prediction by an individual trader might create a rewarding investment opportunity for other traders correcting for the 'wrong' prediction, while in the second case, a 'wrong' prediction in the survey will bias the survey mean. Within a forecasting survey framework, it is harder to neutralize the effect that an off-forecast might have on overall accuracy. In addition, information is often spread around individual forecasters, for example, when it is generated by independent signals, and the aggregation of this information cannot be fully extracted by averaging participants' forecasts (Baron, Mellers, Tetlock, Stone, & Ungar, 2014).

Future research should focus on how to improve both calibration and accuracy of forecasts elicited through prediction markets and surveys. One interesting avenue to explore is to model the starting price in prediction markets as a function of the observable characteristics of the studies or as a function of pre-market survey results. Another interesting topic to explore is weighting or adjusting survey results to improve predictions. A third topic is extending the scope of prediction markets from the ‘binary’ question: ‘will a specific hypothesis successfully replicate?’ to the ‘continuous’ question: ‘which will be the effect size obtained by the replication?’ Empirical evidence about this kind of markets is still lacking (an exception is Forsell et al. 2018), and further research is needed to optimally integrate binary and continuous markets to improve overall accuracy rates.

Literature

1. Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., … Mellers, B. (2017). Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. Management Science, 63(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374
2. Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two Reasons to Make Aggregated Probability Forecasts More Extreme. Decision Analysis. https://doi.org/10.1287/deca.2014.0293
3. Benjamin et al. (2018). Redefine statistical significance. Nature Human Behaviour, 2, (6–10)
4. Brier (1950). Verification of forecasts expressed in terms of probability. Monthly weather review, 78(1), 1–3.
5. Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. Science, 351(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918
6. Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour, 2(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z
7. Cova, F., Strickland, B., Abatista, A. G., Allard, A., Andow, J., Attie, M., …, & Cushman, F. (2018). Estimating the reproducibility of experimental philosophy. https://doi.org/10.1007/s13164-018-0400-9
8. G.Cumming (2008), Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. Perspectives on Psychological Science
9. Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., … Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. Proceedings of the National Academy of Sciences, 112(50), 15343–15347. https://doi.org/10.1073/pnas.1516179112
10. Ebersole et al. (2016). Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology 67 (2016) 68–82
11. Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., … Dreber, A. (2018). Predicting replication outcomes in the Many Labs 2 study. Journal of Economic Psychology. https://doi.org/10.1016/j.joep.2018.10.009
12. Hanson, R. (2003). Combinatorial Information Market Design. Information Systems Frontiers, 5(1), 107–119. https://doi.org/10.1023/A:1022058209073
13. Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. PLOS Medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

14. Ioannidis and Doucouliagos (2013). What's to know about the credibility of empirical economics? Journal of Economic Surveys, 27(5), pp. 997–1004 doi: 10.1111/joes.12032
15. Klein, R. A., et al. (2014). Investigating Variation in Replicability A ''Many Labs'' Replication Project. Social Psychology 2014; Vol. 45(3):142–152 DOI: 10.1027/1864-9335/a000178
16. Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. Advances in Methods and Practices in Psychological Science, 1(4), 443–490. https://doi.org/10.1177/2515245918810225
17. Maniadis et al. (2014). One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. American Economic Review 2014, 104(1): 277–290 http://dx.doi.org/10.1257/aer.104.1.277
18. C. F. Manski, Interpreting the predictions of prediction markets. Econ. Lett. 91, 425–429 (2006). doi:10.1016/j.econlet.2006.01.004
19. Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716
20. Plott & Sunder (1988). Rational Expectations and the Aggregation of Diverse Information in Laboratory Security Markets Econometrica, Vol. 56, No. 5 (Sep., 1988), pp. 1085-1118
21. Schweinsberg et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. Journal of Experimental Social Psychology 66 (2016) 55–67

APPENDIX A

Table A1 provides additional information on the prediction markets performances. It complements the material already described in the manuscript. Refer to the original publications for further details.

Table A1: prediction market performances

	RPP	EERP	ML2	SSRP
N traders	83	97	79	92
Mean N of participants per hypothesis	48	46	44	43
Range participants per hypothesis	22-61	31-68	30-56	18-68
N transactions	2496	2080	2058	1696
Mean N of transactions per hypothesis	109	116	74	81
Range transactions per hypothesis	64-170	72-194	46-108	26-139
Traders' recruitment	OSF and RPP collaboration e-mailing lists	ESA e-mailing list; members of the editorial board of the following econ journals: AER, QJE, RES, ECMA, JPE, EE, JEBO, GEB	RPP collaboration and Open Science Collaboration e-mailing lists	ESA, Society for Judgment and Decision Making, and OSF e-mailing lists; PsychMAP facebook group; Brian Nosek Twitter account
Endowment for trading ($ equivalent)	$100	$50	$50	$50

Table A2 reports the data illustrated in table 2 of the manuscript but focusing only on the 103 studies for which both the prediction markets and the prediction surveys were performed. As Table A2 is built using the same number of underlying studies for both the elicitation methods, it facilitates the comparison between them.

Table A2: Comparison between survey and prediction markets using the same number of studies in both (103)

	RPP	**EERP**	**ML2**	**SSRP**	**Pooled data**
N studies	40	18	24	21	103
Successful replications	15	11	11	13	50
Replication share	37.5%	61.1%	45.8%	61.9%	48.5%
Correct PM (%)	28 (70%)	11 (61%)	18 (75%)	18 (86%)	75 (73%)
Mean beliefs PM	0.556	0.751	0.644	0.634	0.626
Range beliefs PM	0.132 - 0.879	0.588 - 0.937	0.271 - 0.923	0.231 - 0.955	0.132 - 0.955
Mean APE PM	0.431	0.414	0.354	0.303	0.384
Correct Survey (%)	23 (58%)	11 (61%)	16 (67%)	18 (86%)	68 (66%)
Mean beliefs survey	0.546	0.711	0.647	0.605	0.610
Range survey	0.339 - 0.740	0.542 - 0.863	0.327 - 0.887	0.278 - 0.812	0.278 - 0.887
Mean APE Survey	0.485	0.409	0.394	0.348	0.423
Spearman Correlation - PM and Survey beliefs	0.736	0.792	0.947	0.845	0.837