Replication Markets in the Social and Behavioural Sciences

Michael Gordon[1*], Thomas Pfeiffer[1], Domenico Viganola[2] and Yang Liu[3]
[1]Massey University, Auckland, New Zealand
[2]George Mason University, Fairfax, VA
[3]University of California, Santa Cruz, CA
[*]corresponding author: m.b.gordon@massey.ac.nz

Abstract

The credibility of scientific findings is of fundamental importance for future related research. One approach of assessing credibility is to elicit beliefs about the reproducibility of scientific claims from scientists. Four studies recently used surveys and prediction markets to estimate beliefs about replication in systematic large-scale replication projects. The sample sizes in each study were small, which constrained the ability to test a number of hypotheses regarding the performance of prediction markets and surveys. Here, we pooled data from these four studies (n = 103; made available via the R package "PooledMarketR") to assess the performance of surveys and prediction markets. Both average survey responses and prediction market forecasts were highly correlated with replication outcomes (correlations > 0.5). Prediction markets predicted somewhat better than surveys, with lower prediction errors and a higher rate of correct predictions (73% versus 66%). Our results suggest that peer scientists are optimistic, with average beliefs about 10 percentage units higher than the observed replication rates.

1. Background and Summary

The communication of research findings in scientific publications plays a crucial role in the practice of science. However, relatively little is known about how reliable and representative the disseminated pieces of information are. Concerns have been raised about the credibility of published results following John Ioannidis' landmark essay "Why most published findings are false" (Ioannidis, 2005), and the identification of a considerable number of studies that were later shown to be false positives (Ioannidis & Doucouliagos, 2013; Maniadis et al., 2014). In response, a number of large-scale replication projects were initiated in the behavioural and social sciences (Camerer et al., 2016, 2018; Cova et al., 2018; Ebersole et al., 2016; Klein et al., 2014, 2018; Open Science Collaboration, 2015; Schweinsberg et al., 2016) to systematically evaluate a large sample of studies from specific research fields through direct replication. The rates of successful replication in these projects were poor, ranging from 39% to 62%.

Four systematic replication projects were accompanied by prediction markets and surveys aimed at forecasting the replication outcomes before the replications were conducted. The purpose of these prediction market studies was to investigate whether opinions elicited from within research communities are useful predictors of which studies are likely to replicate; and whether prediction markets and surveys are useful mechanisms for eliciting such information from scientists.

In this paper, we analysed a combined data set from four studies that elicited peer beliefs about the replication outcomes of 103 published studies in the social and behaviour sciences. We have made available the data in an R package – 'PooledMarketR'[1]. By pooling the data of the four projects into a single dataset, we substantially increase the statistical power to test the performance of the prediction markets and surveys. In what follows, we provide a Methods section with a brief review of the sources and methodology used in the large-scale replication projects and the prediction market studies, followed by Results and Discussion.

2. Methods

Of the large-scale replications studies conducted over the past decade, the results of four were forecasted by prediction markets and surveys. These large-scale replication studies are:

- Replication Projection: Psychology (RPP) (Open Science Collaboration, 2015)
- Experimental Economics Replication Project (EERP) (Camerer et al., 2016)
- Many Labs 2 (ML2) (Forsell et al., 2018)

[1] https://github.com/MichaelbGordon/PooledMarketR

- Social Science Replication Project (SSRP) (Camerer et al., 2018)

In each study, a set of original studies were selected to be repeated using similar materials and protocols, but usually with larger samples and new participants. Original studies were selected on the basis of a set of pre-defined criteria, including research methodology, specific target journals and time windows. Typically, one key finding of a publication was selected to be replicated with a methodology as close as possible to the original paper. Authors of the original studies were contacted and asked to provide feedback on the replication designs before starting the data collection for the replications.

Statistical power for the replications was typically higher in the replications that in the original studies. RPP and EERP had a statistical power of about 90% to find the original effect size. Following concerns that effect sizes in original studies may be inflated (Ioannidis, 2005; Open Science Collaboration, 2015), therefore increasing the chance of false negatives in replications in the RPP and EERP studies, the power was increased substantially for the SSRP study. This was done by using a 2-stage design, where 90% power was used to detect 75% of the original effect sizes in the first stage and 50% of the original effect size in the second stage. This two-stage approach is further explained below. In the ML2 study, replications were conducted at multiple sites, with greater power.

A binary criterion was used to determined replication success. For the RPP, the EERP, and the SSRP, a replication was deemed successful if it finds a 'significant effect size at 5% in the same direction of the original study'(Cumming, 2008; Open Science Collaboration, 2015); for the ML2, a replication was 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' (Klein et al., 2018). 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 (Klein et al., 2018). Alternative binary and non-binary variables such as effect sizes are reported in the replication studies.

The four prediction markets sought to answer the same question: can we use crowdsourcing to accurately forecast which published studies will replicate? Data from the forecasting projects were easily pooled because the projects shared a similar design. 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. 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 was replicated, and 0 tokens if it was not replicated. 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. 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). An automated market maker implementing a logarithmic market scoring rule was used to determine prices (Hanson, 2003). The prediction markets were open for 2 weeks in the RPP, the ML2, and the SSRP, and for 10 days in the EERP. Detailed 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.

Participants were recruited via blogs, mailing lists and twitter – with the focus on people in academia. Some participants who filled out the survey did not participate in the prediction markets, but the data presented below is restricted to only those 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), so that both the survey and prediction market data is based on the same participants. However, as the survey data are not available for one study of the RPP project, we analyse only those data for which both prediction market and survey data were available.

The rest of this section provides a brief summary of the projects; further details are available in the original publications.

2.1. Dataset 1: Using prediction markets to estimate the reproducibility of scientific research

The study by Dreber et al. (2015) was part of the large scale Replication Project: Psychology (Open Science Collaboration 2015). A subset a set of studies published in the Journal of Personality and Social Psychology, Psychological Science, and Journal of Experimental Psychology were used for eliciting beliefs on the likelihood of successful replication. Dreber et al. ran 41 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 the probability of successful replication. The overall replication rate was 39%. The prediction markets correctly predicted the outcome of the replications 71% of the time, compared with 58% accuracy on the survey.

2.2. Dataset 2: Evaluating replicability of laboratory experiments in economics

Camerer et al. 2016 replicated 18 studies in the field of experimental economics, published in two of the top-5 economic journals (American Economic Review and Quarterly Journal of Economics). The process for selecting the result to be replicated from each study was as follows: (1)

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; (2) if there was more than one equally central result, the result (if any) related to efficiency was picked, as efficiency is central to economics; (3) if several results still remained and they were from different separate experiments, the last experiment (in line with RPP) was chosen; (4) if several results still remained, one of those results was randomly selected for the replication. The fraction of successful replications was 61% and the estimated relative effect sizes in the replicated study were on average 66% that of the original study (i.e., 34% lower, on average). Unlike previous projects (such as RPP), the replication and the forecasts were completed in the same project. Both the markets and the survey correctly categorized 11 studies out of 18 (61%).

2.3. Dataset 3: Predicting replication outcomes in the Many Labs 2 study

Forsell et al. (2018) present the results of the Many Labs 2 study, another on a large replication project led 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 by using large sample sizes across different cultures and labs and requiring replication protocols to be peer-reviewed in advance. The papers were selected by the authors of the Many Labs 2 project, with the aim of assuring diversity and plurality of claims. The realized replication rate for the ML2 project was 46% (11 successful replications out of 24 studies analyzed). The forecasting effort focuses on 24 studies. The prediction markets correctly predicted 75% of the replication outcomes. As a comparison, the survey correctly predicted 67% of replication outcomes.

2.4. Dataset 4: Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015

Dataset 4 comes from a replication project of 21 experimental social science studies published in two general science outlets: Science and Nature (Camerer et al., 2018). The SSRP was specifically designed to address the issue of inflated effect sizes in original studies. There were 3 criteria for selecting studies (presented in descending order): (1) select the first study that reports a significant effect; (2) select the statistically significant result identified as the most important; (3) randomly select a single result in cases of more than one equally central result. In line with previous projects, Camerer et al. 2018 also ran prediction markets and prediction surveys to forecast whether the selected studies will replicate. The design of the SSRP for conducting replications differed from the previous projects in that it was structured in two stages: first, it considered 90% power to detect 75% of the original effect size; if the replication failed, stage 2 started and the data collection kept running until the power of detecting 50% of the original effect size reached 90% (pooling data from stage 1 and stage 2 collection phases). Based on all the data collected, 62% of the 21 studies were successfully replicated. The prediction markets followed 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 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.

3. Results

3.1. Descriptive statistics.

Successful rates of replication ranged from 39% to 62%, with an overall rate of 49% (Table 1). For the prediction markets, we interpreted the final price of each claim as the elicited probability that the claim would replicate. In particular, we interpreted a final price of 0.50 or greater as meaning the market predicting 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 computed 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.

Aggregating the survey using a simple average, the surveys never outperformed 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 were correct 73% of the time (75/103 studies), while the prediction surveys were correct 66% of the time (68/103 studies). The markets had a lower mean absolute error for each of the projects with the exception of EERP where the absolute prediction error is slightly lower for the survey. The Spearman correlation is high between markets and survey beliefs in all the four projects ranging between 0.736 and 0.947.

Prediction markets tend to provide more extreme forecasts with 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.

Table 1: main features of individual projects

	RPP	EERP	ML2	SSRP	Pooled data
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 (1977-2014)	Science, Nature (2010-2015)	
N. studies	40	18	24	21	103
Successful replications	15 (38%)	11 (61.1%)	11 (45.8%)	13 (61.9%)	51 (49%)
Mean beliefs PM	0.556	0.751	0.644	0.634	0.627
Correct PM (%)	28(70%)	11 (61%)	18 (75%)	18 (86%)	76 (73%)
Mean APE PM	0.43	0.414	0.354	0.303	0.383
Mean beliefs survey	0.546	0.711	0.647	0.605	0.61
Correct Survey (%)	23 (58%)	11 (61%)	16 (67%)	18 (86%)	68 (66%)
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
Spearman Correlation – Replication Outcomes and Prediction Market	0.418	0.297	0.755	0.842	0.568
Spearman Correlation – Replication Outcomes and Survey beliefs	0.243	0.516	0.731	0.76	0.557

3.2. 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. 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 ensure that our results are comparable with those reported in previous prediction market publications (Camerer et al., 2016, 2018; Dreber et al., 2015; Forsell et al., 2018) , we also report the non-parametric equivalents.

For all the results reported below, the tests are 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).

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: $APE_{ip}^{pm} = \left|f_{ip} - O_{ip}\right|$ 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 $APE_{ip}^{su} = \left|b_{ip} - O_{ip}\right|$ where b_{ip} is the average belief elicited through the survey for study i in project p.

3.3. Market and Survey Performance

Of the 31 studies that are predicted by the market to not replicate (final prediction market price above 0.5), only 3 eventually replicated, thus for these studies the market is correct more than 90% of the times. On the other hand, out of the 72 studies that are predicted to successfully replicate, 25 did not replicate, with a correct prediction rate of 65%. 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 = 103 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. However while the same information is elicited through the markets and the survey (through the beliefs of the same forecasters), the market aggregates this information better.

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). 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.

When identifying correct predictions (using the binary approach of value above 0.5 indicating a prediction of will replicate), the prediction markets are correct in 8 additional cases with respect to the surveys (75 out of 103 correct forecasts for the markets, 68 out of 103 for the survey). To test if this difference is statistically significant we use a non-parametrical (via Wilcoxon signed-ranks test between the correct predictions of the prediction markets and the correct predictions of the survey) and a parametrical test (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. 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 and on the average beliefs elicited through the surveys. 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 ($p = 0.036$). 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 ($p < 0.001$).

3.4. Analysis of error rates

An additional method of assessing forecasting accuracy is to determine the absolute prediction errors of the forecasts. The average prediction errors of the prediction markets (APE^{pm}, mean = 0.383, median = 0.347, range = [0.045; 0.920], n = 103) 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 ($p < 0.001$).

The accuracy of the forecasts can also be measured in terms of the Brier score, a proper scoring rule scores forecasts against outcomes in a scale between 0 and 1. Higher levels of the Brier score are associated with more inaccurate 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 $B_{ip}^{su} = (b_{ip} - O_{ip})^2$.

The difference between the accuracy rates of the markets and of the survey is less pronounced 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 0.192, 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: $p = 0.205$).

4. Discussion

In this paper, we investigated the forecasting performances obtained by two different procedures to elicit beliefs about replication of scientific studies: prediction markets and surveys. 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, overall, the prediction markets correctly identify which studies successfully replicate and which do not 73% of the times (75/103), 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.

Overall, markets performed better than surveys as a method of aggregating beliefs and providing accurate forecasts. 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. While 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.

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 forecasting binary outcomes (replication success) to continous outcomes (for example effect sizes). 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 the overall accuracy of forecasts.

5. Bibliography

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … 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

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., … Zhou, X. (2018). Estimating the Reproducibility of Experimental Philosophy. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-018-0400-9

Cumming, G. (2008). Replication and p Intervals: P Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science*, *3*(4), 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & 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

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., & 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

Hanson, R. (2003). Combinatorial Information Market Design. *Information Systems Frontiers*, *5*(1), 107–119. https://doi.org/10.1023/A:1022058209073

Ioannidis. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J., & Doucouliagos, C. (2013). What's to Know About the Credibility of Empirical Economics? *Journal of Economic Surveys*, *27*(5), 997–1004. https://doi.org/10.1111/joes.12032

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … 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

Maniadis, Z., Tufano, F., & List, J. A. (2014). One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. *American Economic Review*, *104*(1), 277–290. https://doi.org/10.1257/aer.104.1.277

Manski, C. F. (2006). Interpreting the predictions of prediction markets. *Economics Letters*, *91*(3), 425–429. https://doi.org/10.1016/j.econlet.2006.01.004

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L. L., Diermeier, D., Heinze, J. E., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., du Plessis, C., Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y., … Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. https://doi.org/10.1016/j.jesp.2015.10.001