Predicting replication outcomes in the Many Labs 2 study

Eskil Forsell[1,*], Domenico Viganola[2,*], Thomas Pfeiffer[3], Johan Almenberg[4], Brad Wilson[5], Yiling Chen[6], Brian A. Nosek[7,8], Magnus Johannesson[2] & Anna Dreber[2,9,#]

[1] Spotify Sweden AB, Birger Jarlsgatan 61, SE-113 56 Stockholm, Sweden

[2] Department of Economics, Stockholm School of Economics, Box 6501, SE-113 83 Stockholm, Sweden

[3] New Zealand Institute for Advanced Study, Private Bag 102904, North Shore Mail Centre, Auckland 0745, New Zealand

[4] National Institute of Economic Research, Kungsgatan 12-14, Stockholm, Sweden

[5] Consensus Point, 2323 21st Avenue South, Suite 500, Nashville, TN 37212, USA

[6] John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge MA 02138, USA

[7] Department of Psychology, University of Virginia, Charlottesville, VA 22904, USA

[8] Center for Open Science, Charlottesville, VA 22903, USA

[9] Department of Economics, University of Innsbruck, Innrain 52, 6020 Innsbruck, Austria

* These authors contributed equally to this work.

[#] Corresponding author. E-mail: anna.dreber@hhs.se

Abstract

Understanding and improving reproducibility is crucial for scientific progress. Prediction markets and related methods of eliciting peer beliefs are promising tools to predict replication outcomes. We invited researchers in the field of psychology to judge the replicability of 24 studies replicated in the large scale Many Labs 2 project. We elicited peer beliefs in prediction markets and surveys about two replication success metrics: the probability that the replication yields a statistically significant effect in the original direction ($p<0.001$), and the relative effect size of the replication. The prediction markets correctly predicted 75% of the replication outcomes, and were highly correlated with the replication outcomes. Survey beliefs were also significantly correlated with replication outcomes, but had higher prediction errors. The prediction markets for relative effect sizes attracted little trading and thus did not work well. The survey beliefs about relative effect sizes performed better and were significantly correlated with observed relative effect sizes. These results suggest that replication outcomes can be predicted and that the elicitation of peer beliefs can increase our knowledge about scientific reproducibility and the dynamics of hypothesis testing.

Key words: reproducibility, replications, prediction markets, beliefs.

JEL codes: C92, D80

PsycINFO codes: 226

1. Introduction

For decades, methodologists have been identifying research and publication practices that affect the credibility of published results (Cohen, 1962; Greenwald, 1975; Rosenthal, 1979; Sterling, 1959). Discussion of these issues has surged in the present decade with new theory and evidence that there is indeed cause for concern regarding the validity of published findings due to factors such as publication bias (Lane & Dunlap, 1978; Dwan et al.,2008; Gerber & Malhotra, 2008; Masicampo & Lalande, 2012; Ioannidis et al., 2014; Franco et al. 2014; Rosenthal, 1979; Sterling, 1959), selective reporting of results (Casey et al., 2012; Greenwald, 1975; Humphreys et al., 2013; Simonsohn et al., 2013; Franco et al. 2014), low statistical power (Cohen, 1962; Ioannidis 2005, 2008; Button et al., 2013; Gelman & Carlin, 2014; Maniadis et al., 2014; Nuzzo, 2014), "researcher degrees of freedom" such as p-hacking and data-contingent analysis decisions (Simmons et al., 2011; John et al., 2012; Gelman & Loken, 2013), and, in some cases, outright fraud (Stroebe et al., 2012).

An obvious action in the face of these concerns is to conduct replications to assess the credibility of published findings (Kahneman, 2012; Pashler & Harris, 2012; Bohannon, 2014; Ioannidis & Doucouliagos, 2013; Ioannidis, 2014; Miguel et al., 2014). This inspired several systematic replication projects in the social sciences. In 2015, the "Reproducibility Project: Psychology" (RPP) replicated 100 studies in psychology and only 35 (36%) of the 97 original studies reporting "positive findings" had a significant effect in the original direction in the replication (Open Science Collaboration, 2015).[1] We refer to a statistically significant effect in the original direction as the "statistical significance criterion" for judging replication success for

[1] The remaining 3 original studies did not report "positive findings."

the rest of the paper. In a related replication effort in psychology, the Many Labs projects, the robustness of results from psychological studies across different contexts is investigated (Klein et al., 2014, 2018, Ebersole et al., 2016). A difference between the RPP and the Many Labs projects is that in the RPP each study was only replicated by one lab, whereas in the Many Labs projects several labs replicate each study leading to very large sample sizes and replication power. Based on the statistical significance criterion for replication, 10 out of 13 (77%) studies replicated in Many Labs 1 (Klein et al., 2014), 14 out of 28 (50%) studies replicated in Many Labs 2 (Klein et al., 2018), and 3 out of 10 studies (30%) replicated in Many Labs 3 (Ebersole et al., 2016).[2]

The replication drive in social sciences is not limited to psychology. In economics, a large-scale systematic replication effort was undertaken by Camerer et al. (2016), who replicated 18 studies from two top economics journals in the Experimental Economics Replication Project (EERP). Based on the statistical significance criterion, they found that 11 (61%) out of the 18 studies replicated. The Social Sciences Replication Project (SSRP) replicated 21 systematically selected experimental studies in the social sciences published in Nature and Science between 2010 and 2015, and they found that 13 (62%) of the 21 studies replicated based on the statistical significance criterion (Camerer et al., 2018). Another recent replication project replicated 40 empirical studies in philosophy, and 29 (78%) out of 37[3] original studies reporting “positive findings” replicated based on the statistical significance criterion (Cova et al. 2018).

Given the modest replication rate in the above projects, it is important to understand to what extent the academic community is aware of the limited

[2] Each Many Labs project has a relatively small sample of studies with idiosyncratic inclusion criteria, making it difficult to compare the replication rates across the studies.

[3] The remaining 3 studies were excluded from the analysis because they presented null results.

replicability and is able to anticipate it when interpreting published research findings. Dreber et al. (2015) explored this issue using prediction markets to elicit peer beliefs about replicability. Psychologists traded on the outcomes of 41 of the studies replicated in the RPP. They found that the prediction markets correctly predicted the outcomes of 71% of the replications and that final market prices were significantly correlated with the replication outcomes (Pearson correlation=0.42). Peer beliefs about replication were also elicited in a survey prior to the prediction markets. Peer beliefs from the survey predicted outcomes slightly less well than the prediction markets with 58% of replications correctly predicted (and a Pearson correlation of 0.27). Camerer et al. (2016) used the same procedure to elicit peer beliefs among economists in the EERP. Both the prediction markets and the survey correctly predicted 61% of the replications, but this time the correlation between peer beliefs and replication outcomes was moderately stronger for the survey (Spearman correlation of 0.52 for the survey versus 0.30 for the prediction markets). Peer beliefs using surveys and prediction markets were also collected in the SSRP: both the prediction markets and the survey correctly predicted 86% of the replications, and the correlation between peer beliefs and replications was somewhat stronger for the prediction markets (Spearman correlation 0.84 for the prediction markets versus 0.76 for the survey).

Peer beliefs are not only important for investigating to what extent scientists can predict reproducibility, but they can also be viewed as an additional reproducibility indicator (given that they are correlated with replication outcomes). Furthermore, Dreber et al. (2015) showed how peer beliefs can be used in combination with replication results to estimate the probability that a tested hypothesis is true at different stages of the research process, including the prior of the tested hypothesis.[4]

More work is needed to understand the value of prediction markets and surveys in predicting the outcomes of replication efforts. In this paper, we report the results of prediction markets and surveys on predicting the outcomes of 24 of the studies replicated in the Many Labs 2 project (Klein et al. 2018), where the focus was to replicate 28 new and classic effects across many and diverse samples (>100 samples). We estimated peer beliefs about a successful replication as above (i.e., a significant effect in the same direction as in the original study). In addition to this binary replication outcome we included peer beliefs about relative effect sizes, which can be considered a continuous measure of replication success. We focus on 24 out of these 28 studies as explained in the Appendix, excluding the four original studies that focused on cultural differences in effect sizes between different samples.

In the Many Labs 2 project, 11 out of the 24 studies replicated based on the statistical significance criterion with a strict significance threshold (p < .0001) given the very large samples. The prediction markets successfully predicted 18 (75%) of the 24 replication attempts and the market prices were highly correlated with the replication outcomes (Spearman correlation=0.755). The survey successfully predicted 16 (67%) of the 24 replication attempts, and was also significantly correlated with the replication outcomes (Spearman correlation=0.731). The prediction markets for relative effect sizes did not work well, with little trading and little spread in the predicted effect sizes. The predicted relative effect sizes were moderately correlated with the observed relative effect sizes of the replications (Spearman correlation=0.484), but the size of the effect sizes was substantially

[4] Also based on the RPP data, Johnson et al. (2017) estimate the prior of the original hypotheses tested in the RPP using a different method and find similarly low average priors as Dreber et al. (2015).

overestimated. The effect size prediction markets were more complicated than the binary markets, and participants could choose to focus on the markets for binary outcomes, which probably led to the poor performance. Further work is thus needed to construct well-functioning prediction market for effect sizes. The survey produced average predictions of relative effect sizes that were closer to those observed, and the correlation between survey beliefs and observed relative effect sizes was relatively high (Spearman correlation=0.614).

Our statistical power is limited with just 24 studies investigated.[5] However, this study is part of a cumulative effort to collect prediction market and survey data to assess their performance in predicting the results of scientific studies. We discuss the accumulated data thus far when concluding the paper.

2. Background and design of the study

The idea of prediction markets stems from early work showing that markets can accurately aggregate information (see for example Smith, 1962; Plott & Sunder, 1982; Plott & Sunder, 1988) in combination with work on how to use scoring rules to incentivize individuals to accurately reveal their beliefs about the probabilities of future events (Brier, 1950; McCarthy, 1956; Savage, 1971; Gneiting & Raftery, 2007). Prediction markets have been used for a variety of events such as in sports (Gil & Levitt, 2007; Goel et al., 2008; Borghesi, 2009; Paul & Weinbach, 2009), politics (Forsythe et al., 1992; Bohm & Sonnegård, 1999; Wolfers & Leigh, 2002), and business (Rhode, 2009; Cowgill et al., 2009; O'Leary, 2011). There is substantial

[5] The average correlation between prediction market prices and replication outcomes across the three previous prediction market studies on replications in RPP (Dreber et al. 2015), EERP (Camerer et al. 2016) and SSRP (Camerer et al., 2018) is 0.52. The statistical power in our study to detect a Spearman correlation of 0.52 between prediction market prices and replication outcomes is 75% at a 0.05 significance level (suggestive evidence) and 41% at a 0.005 (statistical significance) level. However, the lack of independence across prediction market predictions and replication outcomes pairs is not taken into account in the power calculation or the Spearman correlation test.

evidence that prediction markets can generate relatively accurate forecasts and in many cases outperform alternative methods (Plott & Chen, 2002; Berg & Rietz, 2003; Debnath et al., 2003; Wolfers & Zitzewitz, 2004; Arrow et al., 2008; Berg et al., 2008a,b; Almenberg et al., 2009; Ledyard et al., 2009; Arnesen & Bergfjord, 2014). A theoretical advantage of prediction markets over individual belief elicitation is that market outcomes do not necessarily correspond to a simple averaging of individual beliefs as information is shared through the trades being made (Plott & Sunder, 1988). The prediction markets we use implement an automated market maker that turns a strictly proper scoring rule for individuals into a mechanism to elicit and aggregate multiple individuals' beliefs (Hanson, 2003, 2007).

The Many Labs 2 project (Klein et al., 2018) studied the replicability of 28 studies published in psychological science and was supported by the Center for Open Science. Table A1 in the Appendix shows a summary of the tested hypotheses as presented to the traders in this study. The project's main goal was to estimate the variation in effect magnitudes across different samples. The effects were divided into two "slates" consisting of 13 and 15 studies respectively. More than 100 teams completed one of the two slates (some did both), and each subject participated in one slate only. Due to the number of teams involved, the slates were administered through a web-based interface, and the 28 studies were chosen in part based on the feasibility of testing them through a web browser. As each slate tested a large number of studies, they were also selected based on the length of the procedure. The replications in the Many Labs 2 project were performed in 2014 and the results were expected in early 2015. Due to the massive scope of the project, the analysis took longer than expected and the results were not published before 2018. Due to this delay, participants in the prediction markets and survey were paid with gift cards in November 2016 according to closing prices in the markets rather than the actual replication outcomes.[6]

The markets were set up in collaboration with Consensus Point, a provider of prediction market research technology. They were of a “winner-takes-all” style as contracts paid a fixed amount of k if a specific event occurred and nothing otherwise. The events over which the contracts were specified were mutually exclusive and exhaustive, guaranteeing that one and only one event would happen. To ensure that there was always a counterpart to trade contracts with, the markets used a market maker posting buy and sell prices for each contract. The market maker implemented by Consensus Point used the logarithmic scoring rule proposed by Hanson (2007).[7]

The market maker ensures that it is always possible to make trades as all trades go through the market maker and the scoring mechanism ensures that a trader making their last trade in the markets have incentives to invest exactly according to their beliefs. Contracts for event i among n mutually exclusive events are priced according to formula (1):

$$p_i = k \cdot \frac{e^{(a_i+S_i)/b}}{\sum_{j=1\dots n} e^{(a_j+S_j)/b}} \tag{1}$$

where p_i is the price of a contract for event i; k is the value of a contract in i if event i occurs; a_i is a constant adjusting the initial price of the contract; S_i is the total number of contracts sold for event i; b is a parameter adjusting how quickly prices move in response to contract purchases.[8] The score for an event i is p_i/k, and equation (1)

[6] All contracts were settled at the closing price, i.e. if a share had a final price of X cents when the market closed, we used a payoff of X cents per share. This applies to both the binary markets and the binned effect size markets. This approach is not incentive compatible, but since we did not expect there to be such a long delay we had not told participants anything about this possibility. We eventually contacted the participants and offered them to settle at closing prices since it was unclear when the Many Labs 2 results would be ready; nobody disagreed.

[7] This scoring rule is mathematically equivalent to the conditional logit model proposed by McFadden (1974) for discrete choice analysis.

[8] Ensuring that trading is always possible means that a subsidy to the traders might be required. The parameter b also determines the maximal amount of this subsidy.

guarantees that the scores sum to 1 and are bounded between 0 and 1. It also ensures that the score movement in response to a contract purchase is constant in terms of the log odds ratio, hence the score movement for a given investment is lower closer to the boundaries. For each of the Many Labs 2 studies, we set up two different kinds of markets. One market type traded contracts for the outcome of the binary measure of replication success and the other traded contracts for the effect size in the replication in relation to the effect size in the original study.

The events specified in the binary market type were that (i) the replication will find an effect in the same direction as in the original study that is significant at the 0.001 level versus (ii) the replication won't find an effect in the same direction as in the original study that is significant at the 0.001 level. We used the significance threshold of p<0.001 due to the large sample sizes and high power used in the Many Labs 2 project.[9]

The events specified in the effect size market type divided the range of possible relative effect sizes into six distinct bins and specified events as: "The effect size in the replication will fall within the range of bin i." Figure 1 shows the division of the bins.

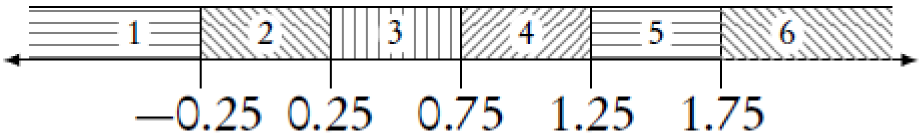

Figure 1: Division of the bins for the effect size in the replication relative to that in the original study. Note that bins 1 and 6 are open-ended.

[9] When we set up the prediction markets and survey, we used p<0.001 as significance threshold since we were informed that this threshold would be used in Many Labs 2. However, the significance threshold used in the published Many Labs 2 paper was p<0.0001 (Klein et al. 2018). Note that whether a p<0.001 or a p<0.0001 threshold is used does not affect the conclusion about replication for the 24 studies included in our results.

The effect sizes were all standardized by conversion to the Cohen's *d* statistic as in the Many Labs 2 research protocol. For one of the studies (study 10 as referenced by Table 1) it was not possible to obtain an original effect size corresponding to the effect being traded on in the binary markets so that study only had a binary market. We thus report results for 24 binary markets and 23 effect size markets.

The market maker offered contracts in all markets at predetermined scores when the markets opened. We set the initial score for both contracts in the binary markets to 0.5 and the initial scores in the effect size markets to 0.225 for contracts for bins 2-5 and 0.05 for contracts for bin 1 and 6. The seemingly low likelihood of a replicated effect ending up in bin 1 or 6 motivated the lower scores for contracts in those bins. Setting initial scores closer to our own beliefs would theoretically have minimized the expected subsidy to traders, but would also have given traders information on our priors. Traders were endowed with $50 expressed as 50.000 "points."[10] The parameter k in the equation above was set to 100 meaning that a contract for an event paid 100 points or $.50 if that event was realized. Traders could invest in any contract or combinations of contracts they liked with the exception of simultaneously holding contracts for both events in the binary markets.[11] They could purchase new contracts or sell currently held ones at any time. A sale returned points to their endowment and could result in a net loss or a gain of points as a result of score changes due to other traders' actions in the time between the purchase and

[10] The trading interface showed this as 10,000 points and hence whenever a trader invested a point the system actually invested five points.

[11] The motivation for this was to remove the option of making a risk free investment by buying a contract for each event, thereby guaranteeing a safe return of k. This trading behavior was still possible in the effect size markets.

sale of the contracts.[12] When the endowment was used up, traders could no longer invest in new contracts and there was no possibility of increasing the endowment by buying more points. To increase incentives for trading, points that were not invested into the markets and were left in the endowment when the markets closed were voided and did not get converted into dollars. The parameter b was set to 1,000. To put this into context, a trader investing her total endowment of 50,000 points in contracts for an event in a binary market where contracts are initially scored at 0.5 would move the score to 0.70. Investing a tenth of their endowment in a similar fashion would move the score to 0.52.

Before the markets opened, the traders were given the following information: our brief summary of the studies in the Many Labs 2 project including the effect size in terms of Cohen's *d* and the *p*-value of the test performed in the original study (see Appendix Table A1); the Many Labs 2 research protocol giving a detailed overview of the replication procedure and more information about the tested hypotheses; and a link to the original paper.[13] The power of the replications was estimated to be very

[12] Participants were permitted to trade in fractional shares. In most instances, a trade is specified with an integer currency amount (e.g., 1,000 points). The amount of points spent will translate into a fractional amount of shares deposited into their trading account. In some cases, a fractional currency amount may also be spent (e.g., if they had sold off a position it would likely deposit fractional currency into their account which would then be spent on their next purchase).

[13] The effect sizes (Cohen's *d*) and *p*-values from the original studies were based on the Many Labs 2 research protocol and the original articles. In some cases there are differences between the effect sizes and *p*-values used in the information to participants, compared to those published in the Many Labs 2 paper (Klein et al. 2018). These are due to slight differences between the Many Labs 2 research protocol and the published Many Labs 2 paper (Klein et al. 2018), and to mistakes. Typically these differences are very small, but for three studies there are important differences in the original *p*-values. For study 18 participants were informed that the *p*-value was 0.004 when it was 0.034; for study 20 participants were informed that the *p*-value was 0.050 when it was <0.001; and for study 24 participants were informed that the *p*-value was <0.001 when it was 0.020. Study 20 is excluded for other reasons as explained in the Appendix, and we do a robustness test to investigate if our results are sensitive towards excluding also study 18 and 24. The original effect sizes and *p*-values we report in Tables 1 and 2 are taken from the effect sizes in the Many Labs 2 paper and the test-statistics for the original results in the Many Labs 2 paper (Klein et al. 2018). In the Many Labs 2 paper they also report effect sizes as Cohen's *q* rather than Cohen's *d* for the two studies (study 8 and study 28) that test for differences in correlation coefficients (Klein et al. 2018). We do the same in Tables 1 and 2 and when we estimate the observed relative effect sizes for these two studies, to make sure that effect sizes are estimated in the same way for the original result and the replication result (although we used

close to 100% as the sample size in each study was expected to be over 4,500 participants. Traders were told it was 99% for all effects. Before being allowed to trade in the markets or view current scores, traders had to answer a pre-market survey consisting of three questions for each of the hypotheses tested in Many Labs 2; (1) “How likely do you think it is that this hypothesis will be replicated (on a scale from 0% to 100%)?” (2) “How large do you think the standardized effect size (in terms of Cohen’s *d*) from the replication will be, relative to that in the original paper (on a scale from -50% to 200%)?”, and (3) “How well do you know this topic? (Not at all; Slightly; Moderately; Very well; Extremely well.)”.

The reason for only letting traders who had answered the survey trade was that this enabled us to keep a balanced sample of traders when comparing the answers from the survey to the results of the markets. In contrast to trading in the markets, survey answers were not incentivized, and traders did not get feedback on the survey responses of others. Traders were recruited from the Replication Project: Psychology and Open Science Collaboration e-mailing lists predominantly consisting of psychologists interested in replications. In order to ensure participation of traders with relevant background knowledge, we specifically asked for them to be currently studying for or having already completed a Ph.D. in psychology. We had no way to enforce this, but as both mailing lists are directed at academics in the field, it is unlikely that anything but a very minor number of traders not fulfilling this criterion was included.

The markets were open during two weeks in the fall of 2014. 107 people responded to our invitation and expressed an interest in participating. Out of these, 91 finished the survey before the markets closed, and out of these, 78 made at least

Cohen’s *d* to describe effect sizes for all the studies in the information to participants in the prediction markets and survey).

one trade in any market. This final sample of 78 is referred to simply as the traders and the results presented below are for this sample.[14] Most of these traders (69/78) invested more than 90% of their cash into forecasts, suggesting that most participants understood that final cash holdings were not redeemed.

3. Results

Here, we present the results for the binary markets and the effect size markets. For all results reported here, we interpret p<0.005 as “statistically significant” and p<0.05 as “suggestive evidence” in line with the recent recommendation of Benjamin et al. (2018). Note that survey beliefs and prediction market beliefs are not independent from each other or over replication studies because they are elicited from the same set of subjects in all cases (i.e. the same subjects participate in both the survey and the prediction markets). In testing for significant differences between survey beliefs and prediction market beliefs we therefore use the Wilcoxon signed-ranks test based on paired data.

3.1 Binary Markets

[14] The survey responses of the 13 subjects that completed the survey but did not trade in the prediction markets did not differ significantly from the subjects that actively participated in the markets in terms of the average predicted replication probability (0.628 vs 0.647, Mann-Whitney U-test, *p*=0.519). However, there is suggestive evidence for a difference in the average predicted relative effect size (0.491 vs 0.396, Mann-Whitney U-test, *p*=0.015); or and the average knowledge of the topics (1.80 vs 2.02, computed on an experience scale ranging from 1 to 5, Mann-Whitney U-test, *p*=0.017). If we estimate the survey responses for all the 91 subjects that completed the survey, the average predicted replication probability is 0.644 instead of 0.647 for the n=78 sample that participated in the markets, and the average predicted relative effect size is 0.409 for the n=91 sample and 0.396 for the N=78 sample; so these means are only marginally affected if the n=91 sample is used. The Spearman correlation between the predicted replication probabilities for the 24 studies is 0.998 for the n=91 and the n=78 samples, and the corresponding Spearman correlation for the predicted relative effect sizes is 0.997.

Trading was high in the binary markets.[15] All traders traded in at least one binary market, and 12 traders traded in all markets (mean: 16, median: 15). The lowest number of traders being active in a single binary market was 30, and the highest number was 56 (mean: 44, median: 44). The number of traders in a market is only a lower bound on the number of transactions; accordingly, the number of transactions in a single binary market ranged between 46 and 108 (mean: 73, median: 71).

Table 1. Market and survey beliefs about replication probability

Study	Survey beliefs (weighted)	Market beliefs	Original *p*-value	Replication *p*-value	Replicated (Yes/No)
Study 2: Kay et al., (2014)	0.327 (0.331)	0.291	0.050	0.35	0
Study 3: Alter et al., (2007)	0.422 (0.449)	0.328	0.051	0.43	0
Study 4: Graham, Haidt & Nosek (2009)	0.765 (0.784)	0.904	$p < 0.001$	$p < 0.001$	1
Study 5: Rottenstreich & Hsee (2001)	0.517 (0.543)	0.524	0.027	0.002	0
Study 6: Bauer et al.(2012)	0.673 (0.682)	0.767	$p < 0.001$	$p < 0.001$	1
Study 8: Inbar et al., (2009)	0.529 (0.585)	0.564	0.034	0.02	0
Study 9: Critcher & Gilovich (2008)	0.459 (0.514)	0.404	0.035	0.09	0
Study 10: Van Lange et al., (1997)	0.573 (0.602)	0.402	0.008	0.18	0
Study 11: Hauser et al., (2007) st. 1	0.852 (0.868)	0.897	$p < 0.001$	$p < 0.001$	1
Study 12: Anderson et al., (2012)	0.705 (0.726)	0.702	0.003	0.08	0
Study 13: Ross, Greene & House (1977) st.1	0.810 (0.834)	0.885	$p < 0.001$	$p < 0.001$	1
Study 14: Ross, Greene & House (1977) st. 2	0.795 (0.817)	0.807	$p < 0.001$	$p < 0.001$	1

[15] The numbers below on trading activities refer to all the 28 markets, and not the 24 studies we can compare to replication outcomes. See the Appendix for the motivation for excluding 4 studies from our results.

Study 15: Giessner & Schubert (2007)	0.507 (0.509)	0.392	0.032	0.16	0
Study 16: Tversky & Kahneman (1981)	0.887 (0.902)	0.923	$p < 0.001$	$p < 0.001$	1
Study 17: Hauser et al., (2007) st. 2	0.683 (0.705)	0.674	$p < 0.001$	$p < 0.001$	1
Study 18: Risen & Gilovich (2008)	0.731 (0.736)	0.742	0.034	$p < 0.001$	1
Study 21: Hsee (1998)	0.741 (0.748)	0.761	0.002	$p < 0.001$	1
Study 22: Gray & Wegner (2009)	0.815 (0.826)	0.820	0.001	$p < 0.001$	1
Study 23: Zhong & Liljenquist (2006)	0.416 (0.385)	0.271	0.014	0.91	0
Study 24: Schwarz, Strack & Mai (1991)	0.723 (0.770)	0.721	0.020	0.002	0
Study 25: Shafir (1993)	0.557 (0.594)	0.607	0.013	$p < 0.001$	0
Study 26: Zaval et al., (2014)	0.447 (0.468)	0.432	0.033	0.27	0
Study 27: Knobe (2003)	0.781 (0.818)	0.798	$p < 0.001$	$p < 0.001$	1
Study 28: Tversky & Gati (1978)	0.804 (0.822)	0.831	0.003	0.55	0

Note: Studies 1, 7, 19, and 20 were excluded; refer to the appendix of the paper.

Table 1 summarizes the main market outcomes for each of the 24 studies. Market beliefs about replication ranged between 0.923 and 0.271 for the 24 studies (mean: 0.644, median: 0.712). The survey beliefs per study ranged between 0.887 and 0.327 (mean: 0.647, median: 0.694). The average belief was very similar for the markets and the survey (Wilcoxon signed-ranks test, p=0.663), and the correlation between market and survey beliefs was very high (Spearman correlation=0.947; $p<0.0001$) (Appendix Figure A1). This correlation was somewhat higher than the correlations of 0.78, 0.79 and 0.85 found in Dreber et al. (2015), Camerer et al. (2016) and Camerer et al. (2018) respectively. The range in the predictions, between the highest and the lowest beliefs, was somewhat higher for the markets than for the survey: 0.652 for the markets versus 0.561 for the survey. The experience question in the survey makes it possible to also construct a weighted survey response of peer beliefs as done by Dreber et al. (2015). However, the results for the weighted survey was very close to the survey without weighting responses with a Spearman correlation of 0.995 between the two for the binary outcomes and 0.980 for the relative effect sizes.[16] We therefore do not further discuss the weighted survey results below, but the weighted survey results are reported in Tables 1 and 2 for completeness.

The peer beliefs can be compared to the replication outcomes in Many Labs 2. In total 11 (46%) of the 24 studies replicated. Both the prediction markets and the survey overestimated the average replication rate, and there was suggestive evidence for this overestimation[17] (Wilcoxon signed-ranks test; *p*=0.018 for market beliefs versus replication outcomes and *p*=0.016 for survey beliefs versus replication outcomes).

One way of estimating how well beliefs predict outcomes for this binary replication outcome is to estimate the fraction of correct predictions with beliefs above 50% interpreted as predicting a successful replication and beliefs below 50% interpreted as predicting a failed replication. This measure is most suitable to evaluate predictions when we can expect a replication rate relatively close to 50% and a wide range of predictions.[18] The market correctly predicted 18/24 (75%) of the replications, which provided suggestive evidence of a correct prediction rate over

[16] The prediction in terms of predicting a successful or failed replication, i.e. a predicted probability above or below 50%, only differed for one study. For study 9 the survey predicted a 46% probability of replication and the weighted survey predicted a 52% probability of replication; this study did not replicate.

[17] There was also a tendency to overestimate the replication rate in the RPP studies explored by Dreber et al. (2015) as well as in the EERP (Camerer et al. 2016), but not in the SSRP (Camerer et al. 2018).

[18] The fraction of correct predictions is a less suitable measure if the replication rate is very high or very low. The fraction of correct predictions can be high in such situations without being able to discriminate between studies replicating and not replicating. Therefore it is important to also use other measures of prediction accuracy as the correlation between prediction market beliefs/survey beliefs and replication outcomes; and to compare the levels of beliefs with the observed replication rate and average effect size.

50% (one-sample binomial test, *p*=0.023).[19] The survey correctly predicted 16/24 (67%) of the replications, which was not significantly different from 50% (one-sample binomial test, *p*=0.152) and also not different from 18/24 correct prediction rate for the market (Wilcoxon signed-ranks test using Pratt's method to account for zero values, *p*=0.157).

In Figure 2 we show the relationship between the market beliefs and the survey beliefs and the replication outcomes. The correlation was quite high between market beliefs and replication outcomes (Spearman correlation=0.755, p<0.0001) and between survey beliefs and replication outcomes (Spearman correlation coefficient=0.731, p<0.0001). To test if the market beliefs predict significantly better than the survey belief we compared the absolute prediction error between the two. The absolute prediction error is the absolute difference between the replication outcome (either 1 if it replicated or 0 if it did not replicate) and the prediction market (survey) beliefs, i.e. the difference between the outcome and the predicted probability. The prediction error was significantly lower for prediction market beliefs (mean absolute prediction error for market beliefs =0.354 and for survey beliefs =0.394; Wilcoxon signed-ranks test, *p*=0.003).

[19] A 50% correct prediction rate would be the expected rate due to chance alone.

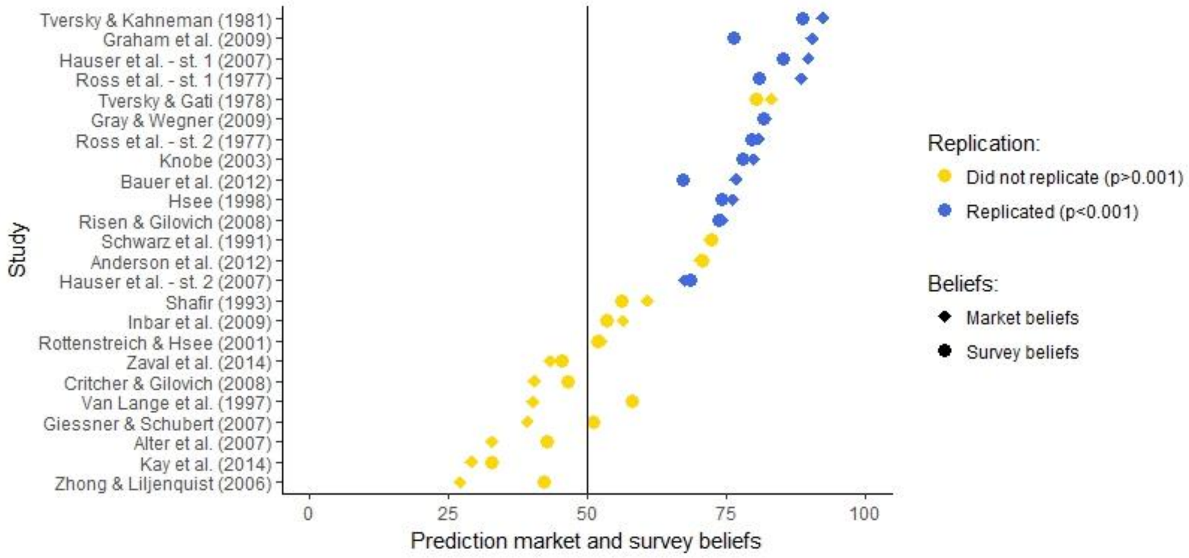

FIGURE 2: Prediction market and survey beliefs for the binary markets (replication probability). The figure shows the prediction market beliefs and the survey beliefs of the replication probability. The replication studies are ranked in terms of prediction market beliefs on the y-axis. The prediction market beliefs (Spearman correlation coefficient 0.755, *p*<0.0001) and the survey beliefs (Spearman correlation coefficient 0.731, *p*<0.0001) are significantly correlated with a successful replication.

The information about the *p*-value of the original study given to the participants in the prediction markets was incorrect for studies 18 and 24 (see footnote 13). We therefore carried out a robustness test excluding studies 18 and 24 (so that the sample of studies is n=22). Excluding these two studies, the average belief was 0.636 for the markets and 0.639 for the survey. Out of these 22 studies, 10 (45%) replicated. There was suggestive evidence that the market and survey beliefs overestimated the replication rate (Wilcoxon signed-ranks test; *p*=0.023 for market beliefs versus replication outcomes and *p*=0.021 for survey beliefs versus replication outcomes). The market correctly predicted 17/22 (77%) of the replications, which provided suggestive evidence of a difference to 50% (one-sample binomial test, *p*=0.017). The survey correctly predicted 15/22 (68%) of the replications, which is not significantly different from 50% (one-sample binomial test, *p*=0.134). The Spearman correlation between beliefs and replication outcomes was 0.763 (p<0.0001) for the market beliefs and 0.748 (p<0.0001) for the survey beliefs. The prediction error was significantly lower for the market beliefs than for the survey beliefs (mean absolute prediction error for market beliefs=0.342 and for survey beliefs=0.385; Wilcoxon signed-ranks test, *p*=0.004). The results were thus very similar in this robustness test.

A potentially important determinant of whether an original study will replicate or not is the *p*-value of the original study. The RPP (Open Science Collaboration, 2015) found a Spearman correlation between the original *p*-value and the replication rate of -0.33, the EERP (Camerer et al., 2016) found a Spearman correlation of -0.57, and the SSRP (Camerer et al. 2018) found a Spearman correlation of -0.40. We estimated this correlation also for the 24 Many Labs 2 studies included in our study and found a significant Spearman correlation of -0.755 (p<0.0001). The original *p*-values are also highly correlated with the market (Spearman correlation=-0.781, p<0.0001) and survey beliefs (Spearman correlation=-0.768, p<0.0001) suggesting that they might be used for forming beliefs about replication. However, an alternative possibility is that there is a third factor, such as the plausibility of the original result, that predicts both observed *p*-values and the survey beliefs and market prices. With either explanation, these findings are consistent with the claim that reproducibility of significant new findings could be improved by using a lower threshold of <0.005 for statistical significance for new findings (Benjamin et al., 2018). Of the 24 original studies from Many Labs 2 included in our results, 12 had a *p*-value <0.005 and 12 had a *p*-value >0.005 (see Table 1). For the 12 studies with an original p<0.005, 10 (83%) replicated. For the 12 studies with an original p>0.005, only 1 (8%) replicated. Further work is needed to test if prediction markets outperform predictions based only on the initial *p*-value, to test if the market also aggregates other information important for reproducibility.

3.2 Effect size Markets

The trading pattern in the effect size markets differs from that in the binary markets in some important ways.[20] Eighteen of the 78 traders never traded in an effect size market and 10 traded in all markets (mean: 9, median: 6). The lower trading intensity compared to the binary markets also occurred in the number of traders per market, the lowest number was 16 and the highest 36 (mean: 26, median: 27). Similarly, the transactions in a single effect size market ranged from 25 to 62 (mean: 42, median: 42).

Table 2. Market and survey beliefs about replication relative effect sizes.

Study	Survey beliefs (weighted)	Market beliefs	Original effect size	Replication effect size	Replication relative effect size
Study 2: Kay et al., (2014)	0.073 (0.092)	0.580	0.49	-0.02	-0.04
Study 3: Alter et al., (2007)	0.199 (0.214)	0.513	0.64	-0.03	-0.05
Study 4: Graham, Haidt & Nosek (2009)	0.506 (0.532)	0.683	0.52	0.29	0.56
Study 5: Rottenstreich & Hsee (2001)	0.258 (0.298)	0.634	0.74	-0.08	-0.11
Study 6: Bauer et al.(2012)	0.357 (0.380)	0.625	0.87	0.12	0.14
Study 8: Inbar et al., (2009)	0.295 (0.332)	0.668	0.70	0.05	0.07
Study 9: Critcher & Gilovich (2008)	0.277 (0.372)	0.601	0.30	0.04	0.13
Study 11: Hauser et al., (2007) st. 1	0.488 (0.495)	0.781	2.50	1.35	0.54
Study 12: Anderson et al., (2012)	0.459 (0.476)	0.666	0.57	-0.04	-0.07
Study 13: Ross, Greene & House (1977) st.1	0.546 (0.560)	0.713	0.99	1.18	1.19

[20] The numbers below on trading activities refer to all the 27 effect size markets, and not the 23 studies we can compare to replication outcomes on relative effect sizes.

Study 14: Ross, Greene & House (1977) st. 2	0.493 (0.514)	0.687	0.79	0.95	1.19
Study 15: Giessner & Schubert (2007)	0.280 (0.295)	0.575	0.55	0.03	0.05
Study 16: Tversky & Kahneman (1981)	0.597 (0.641)	0.733	1.08	0.40	0.37
Study 17: Hauser et al., (2007) st. 2	0.521 (0.543)	0.737	0.34	0.25	0.74
Study 18: Risen & Gilovich (2008)	0.473 (0.473)	0.723	0.39	0.18	0.46
Study 21: Hsee (1998)	0.462 (0.474)	0.699	0.69	0.78	1.13
Study 22: Gray & Wegner (2009)	0.556 (0.560)	0.721	0.81	0.95	1.19
Study 23: Zhong & Liljenquist (2006)	0.185 (0.182)	0.503	1.02	0	0.00
Study 24: Schwarz, Strack & Mai (1991)	0.480 (0.530)	0.699	0.48	-0.07	-0.15
Study 25: Shafir (1993)	0.385 (0.408)	0.716	0.35	-0.13	-0.37
Study 26: Zaval et al., (2014)	0.273 (0.281)	0.588	0.31	-0.03	-0.10
Study 27: Knobe (2003)	0.453 (0.452)	0.799	1.45	1.75	1.21
Study 28: Tversky & Gati (1978)	0.487 (0.521)	0.790	0.48	0.01	0.02

Note: Study 10 was not part of the effect size markets, and studies 1, 7, 19, and 20 were excluded; see the appendix of the paper. All effect sizes presented are in Cohen's *d* units, except study 8 and 24 that are in Cohen's *q* units following the Many Labs paper (Klein et al. 2018).

The peer beliefs in the effect size markets are summarized in Table 2, and the results for each of the 6 bins on the effect size markets are reported in Appendix Table A2.[21] The range of predicted relative effect sizes based on market beliefs was 0.503 to 0.799 for the 23 studies (mean: 0.671, median: 0.687). This range appears narrow given the potentially wide range of outcomes for the relative effect size. Further, the scores for the open-ended bins 1 and 6 tied to events that may have been perceived as unlikely did not move far from the initial scores of 0.05 (the average final scores were 0.055 and 0.040 respectively). The predicted relative effect

[21] The "Market beliefs" column represents the predicted effect size from the prediction market. These were calculated by taking the scores multiplied by the midpoint of each bin. For the open ended bins (1 and 6) -0.5 and 2 were used as midpoints, respectively.

sizes per study based on survey beliefs ranged between 0.073 and 0.597 (mean: 0.396, median: 0.459).[22] There was a substantial difference in the market and survey beliefs for effect sizes, and this difference was significant (Wilcoxon signed-ranks test, p<0.0001). The Spearman correlation of 0.760 (p<0.0001) between the market and survey beliefs was lower than for the binary outcome measure of replication (Appendix Figure A2).

These observations indicate that the score movements in the effect size markets were insufficient to obtain reasonable predictions. The more complex structure of the effects size markets, with contracts for multiple events divided into bins, is likely one important factor behind the reduced trading activity.

With a fixed endowment the traders may prefer to spend their points in the simpler binary markets.[23] The plausibility of a relative effect size may also be harder to assess than the simple binary measure, driving traders to the binary markets.

The peer beliefs for relative effect sizes can be compared to the replication outcomes in Many Labs 2. The average relative replication effect size in Many Labs 2 was 0.353. The market beliefs significantly overestimated the observed relative effect sizes (Wilcoxon signed-ranks test; *p*=0.003), but the survey beliefs did not differ significantly from the observed relative effect sizes (Wilcoxon signed-ranks test, *p*=0.445).

[22] There is a potential issue in comparing the predicted relative effect sizes between the survey and the markets. In the survey subjects directly respond about their predicted relative effect size whereas for the markets bins with different effect size intervals are used to construct the average predicted relative effect size for each market using the midpoint of each bin. To test the importance of this we tested converting the survey responses to bins and then constructed the survey measure of predicted relative effect sizes per study in the same way as for the markets. The Spearman correlation between this alternative survey measure and the survey measure used in the main analyses is 0.984 and the mean predicted relative effect size with the survey using this measure is 0.385 compared to 0.396 for our measure in Table 2. Using this alternative method of estimating the predicted relative effect sizes from the survey thus yields similar results.

[23] Trading in the two market types was related however, the Spearman correlation between a study's predicted relative effect size and the market beliefs in the binary markets was 0.75 (*p* <0.0001).

In Figure 3 we show the relationship between the observed relative effect sizes, the market beliefs, and the survey beliefs. The correlation was positive and moderately high between market beliefs and observed relative effect sizes (Spearman correlation=0.484, *p*=0.019). The correlation was higher between survey beliefs and observed relative effect sizes, and this correlation was significant (Spearman correlation coefficient=0.614, *p*=0.002). To test if the survey beliefs predicted significantly better than the market belief we compared the absolute prediction error between the two, and there was suggestive evidence for a lower prediction error for survey beliefs (mean absolute prediction error for market beliefs=0.517 and for survey beliefs=0.366; Wilcoxon signed-ranks test, *p*=0.007).

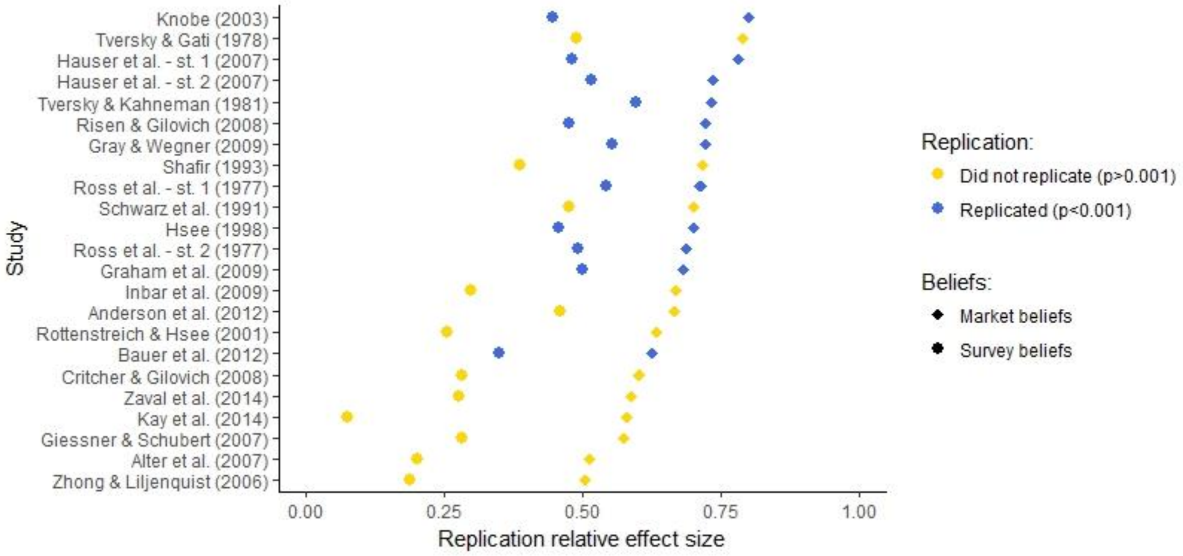

FIGURE 3: Prediction market and survey beliefs for the relative effect size markets. The figure shows the prediction market beliefs and the survey beliefs of the relative replication effect size. The replication studies are ranked in terms of prediction market beliefs on the y-axis. The Spearman correlation coefficient between market beliefs and observed relative replication effect sizes is 0.484 (*p*=0.019) and the Spearman correlation between survey beliefs and observed relative replication effect sizes is 0.622 (*p*=0.001).

We also carried out the same robustness test as for the binary replication outcomes above, excluding studies 18 and 24. Excluding these two studies, the average predicted effect size was 0.667 for the markets and 0.388 for the survey. The observed mean relative replication effect size was 0.349 for these 21 studies. We found suggestive evidence that the predicted relative effect size was higher than the observed for the market beliefs (Wilcoxon signed-ranks test, *p*=0.006), but not for the survey beliefs (Wilcoxon signed-ranks test, *p*=0.585). The correlation between predicted relative effect sizes and the observed relative effect sizes was 0.544 (*p*=0.011) for the market beliefs and 0.686 (*p*=0.001) for the survey beliefs. There was suggestive evidence of a lower absolute prediction error for the survey than for the market (mean absolute prediction error for market beliefs=0.514 and for survey beliefs =0.370; Wilcoxon signed-ranks test, *p*=0.018). This robustness test thus gave similar results as the analyses for the 23 studies above.

4. Concluding Remarks

The prediction markets were quite successful in predicting whether 24 studies included in Many Labs 2 would replicate or not. The markets correctly predicted 75% of the replications, and the Spearman correlation was 0.755 between market beliefs and replication outcomes. However, with only n=24 the uncertainty around these point estimates are large. The point estimates for the prediction markets is somewhat higher than in Dreber et al. (2015) for the RPP replications with 71% correctly predicted and Camerer et al. (2016) for the EERP replications with 61% correctly predicted; but lower than in Camerer et al. (2018) for the SSRP replications with 86% correctly predicted. Based on statistical properties of the replication data, one would expect a higher prediction accuracy for the Many Labs 2 studies than for the RPP and EERP because the replications had higher power and used a lower threshold for significance. This reduced both the false positive risk and the false negative risk of the replications based on statistical significance. Also the SSRP had high powered replications, which may also have contributed to the high prediction accuracy in that study. Pooling the results from all four prediction market studies yields a 73% (76/104) correct prediction rate.

The point estimate for the correct prediction rate for the survey was somewhat lower than for the markets in this study, with a 67% correct prediction rate of the binary replication outcomes. Pooling the results across all four studies for the survey yield a 66% (68/103) correct prediction rate.[24] These results point towards the market predicting replication outcomes somewhat better than the survey, which is also consistent with the significantly lower prediction error for the markets compared to the survey in our data for Many Labs 2. However, the average correlation between peer beliefs and replication outcomes across the four studies is similar between the prediction markets and the survey with an average correlation of 0.58 for the four prediction market studies and 0.57 for the four survey studies. More data are therefore needed to draw firm conclusions about whether prediction markets outperform surveys in this context. We also found suggestive evidence that the market beliefs and the survey beliefs overestimated the replication rate. This tendency was also observed in the Dreber et al. (2015) and the Camerer et al. (2016) studies, but not in the Camerer et al. (2018) study. Taken together this suggests that scientists on average somewhat overestimate the reproducibility of published findings.

The prediction markets worked less well for predicting relative effect sizes, while the survey performed better. The poor performance of the effect size markets may be due to the increased complexity of the multiple intervals, which led to little

[24] The number of observations is one study less for the survey as data was missing for the survey for one study in Dreber et al. (2015).

trading and left the predicted relative effect sizes close to the initial value of 0.75. Participants may also have preferred to trade on the simpler binary markets. In future studies, it may be better to completely separate the binary and effect size markets to encourage more trading on the effect size markets. Also, better instruction may reduce the complexity of effect size markets for traders. For example, letting traders become familiar with the market structure with hypothetical examples might help them understand the relationship between current market prices and the expected effect size ("Market beliefs" column of Table 2). Further, trading might be increased with a simpler market structure with fewer bins or with a continuous market for the effect size as in markets for voting shares. A challenge with this solution is how to deal with the possibility of negative effect sizes and how to avoid negative prices on the markets and negative payments to participants. More work is thus needed on this topic.

Acknowledgments: We are grateful for generous financial support from the Jan Wallander and Tom Hedelius Foundation (Svenska Handelsbanken Forskningsstiftelser), the Knut and Alice Wallenberg Foundation (AD is a Wallenberg Academy Fellows), Marsden Fund Grants 16-UOA-190 and 17-MAU-133 and the Swedish Foundation for Humanities and Social Sciences.

References

Almenberg, J., Kittlitz, K., & Pfeiffer, T. (2009). An experiment on prediction markets in science. *PLoS ONE*, 4(12), e8500.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R.N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, 136(4), 569.

Anderson, C., Kraus, M.W., Galinsky, A. D., & Keltner, D. (2012). The local ladder effect social status and subjective well-being. *Psychological Science*, 23(7), 764–771.

Arnesen, S., & Bergfjord, O. (2014). Prediction markets vs polls – an examination of accuracy for the 2008 and 2012 elections. *The Journal of Prediction Markets*, 8(3), 24–33.

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., Levmore, S., Litan, R., Milgrom, P., Nelson, F. D., & others (2008). The promise of prediction markets. *Science*, 320(5878), 877.

Bauer, M. A., Wilkie, J. E., Kim, J. K., & Bodenhausen, G. V. (2012). Cuing consumerism Situational Materialism Undermines Personal and Social Well-Being. *Psychological Science*, 23(5), 517–523.

Benjamin, D, et al. (2018). Redefine statistical significance. *Nature Human Behaviour* 2, 6-10.

Berg, J., Forsythe, R., Nelson, F., & Rietz, T. (2008a). Results from a dozen years of election futures markets research. Handbook of Experimental Economics Results, 1, 742–751.

Berg, J. E., Nelson, F. D., & Rietz, T. A. (2008b). Prediction market accuracy in the long run. *International Journal of Forecasting*, 24(2), 285–300.

Berg, J. E., & Rietz, T. A. (2003). Prediction Markets as Decision Support Systems. *Information Systems Frontiers*, 5(1), 79–93.

Bohannon, J. (2014). Replication effort provokes praise—and ‘bullying’ charges. *Science*, 344(6186), 788–789.

Bohm, P., & Sonnegård, J. (1999). Political stock markets and unreliable polls. *The Scandinavian Journal of Economics*, 101(2), 205–222.

Borghesi, R. (2009). An Examination of Prediction Market Efficiency: Nba Contracts on Tradesports. *The Journal of Prediction Markets*, 3(2), 63–77.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. Monthly weather review, 78(1), 1–3.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376.

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.

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., Isaksson, S. Manfredi, D., Rose, J., Wagenmakers, E-J., & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, published online.

Casey, K., Glennerster, R., & Miguel, E. (2012). Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan*. *The Quarterly Journal of Economics*, 127(4), 1755–1812.

Cova, Florian, Brent Strickland, Angela G Abatista, Aurélien Allard, James Andow, Mario Attie, James Beebe, et al. (2018). Estimating the Reproducibility of Experimental Philosophy. PsyArXiv. doi:10.17605/OSF.IO/SXDAH.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.

Cowgill, B., Wolfers, J., & Zitzewitz, E. (2009). Using Prediction Markets to Track Information Flows: Evidence from Google. In *AMMA*, (p. 3).

Critcher, C.R., & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making*, 21(3), 241–251.

Debnath, S., Pennock, D. M., Giles, C. L.,& Lawrence, S. (2003). Information incorporation in online in-game sports betting markets. In Proceedings of the 4th ACM conference on Electronic commerce, (pp. 258–259). New York, NY, USA: ACM.

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.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P. J., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J. P. A., Simes, J., & Williamson, P. R. (2008). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. *PLoS ONE*, 3(8), e3081.

Ebersole, C. R. et al. Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology* 67, 68–82 (2016).

Franco, Malhotra & Simonovits (2014) Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: unlocking the file drawer. *Science*. 2014; 345(6203):1502–1505.

Forsythe, R., Nelson, F., Neumann, G. R., & Wright, J. (1992). Anatomy of an experimental political stock market. *The American Economic Review*, 85(5), 1142–1161.

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6), 641–651.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no ‘fishing expedition’ or ‘p-hacking’ and the research hypothesis was posited ahead of time. Downloaded January 30, 2014.

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*.

Giessner, S. R., & Schubert, T.W. (2007). High in the hierarchy: How vertical location and judgments of leaders’ power are interrelated. *Organizational Behavior and Human Decision Processes*, 104(1), 30–44.

Gil, R. G. R., & Levitt, S. D. (2007). Testing the efficiency of markets in the 2002 world cup. *The Journal of Prediction Markets*, 1(3), 255–270.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Goel, S., Pennock, D., Reeves, D. M., & Yu, C. (2008). Yoopick: A Combinatorial Sports Prediction Market. In *AAAI*, (pp. 1880–1881).

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5), 1029.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of personality and social psychology*, 96(3), 505.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82, 1–20.

Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers*, 5(1), 107–119.

Hanson, R. (2007). Logarithmic market scoring rules for modular combinatorial information aggregation. *The Journal of Prediction Markets*, 1(1), 3–15.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & language*, 22(1), 1–21.

Hsee, C. K. (1998). Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, 11.

Huang, Y., Tse, C.-S.,&Cho, K.W. (2014). Living in the north is not necessarily favorable: Different metaphoric associations between cardinal direction and valence in Hong Kong and in the United States. *European Journal of Social Psychology*, 44(4), 360–369.

Humphreys, M., Sierra, R. S. d. l., &Windt, P. v. d. (2013). Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration. *Political Analysis*, 21(1), 1–20.

Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion*, 9(3), 435.

Ioannidis, J., & Doucouliagos, C. (2013). What's to know about the credibility of empirical economics? *Journal of Economic Surveys*, 27(5), 997–1004.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in cognitive sciences*, 18(5), 235–241.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS* Med, 2(8), e124.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5), 640–648.

Ioannidis, J. P. A. (2014). How to Make More Published Research True. *PLoS* Med, 11(10), e1001747.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5), 524–532.

Johnson Valen. E, Richard D. Payne, Tianying Wang, Alex Asher, and Mandal Soutrik. 2017. On the reproducibility of psychological science. *Journal of the American Statistical Association*, in press.

Kahneman, D. (2012). A proposal to deal with questions about priming effects.

Kay, A. C., Laurin, K., Fitzsimons, G. M., & Landau, M. J. (2014). A functional basis for structure-seeking: Exposure to structure promotes willingness to engage in motivated action. *Journal of Experimental Psychology*: General, 143(2), 486.

Klein, R. A. et al. Investigating variation in replicability: a ‘many labs’ replication project.” *Social Psychology* 45, 142–152 (2014).

Klein, R. A. et al. Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, in-principle accepted (2018).

Knobe, J. (2003). Intentional action and side effects in ordinary language. Analysis, 63(279), 190–194.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, 31(2), 107–112.

Ledyard, J., Hanson, R., & Ishikida, T. (2009). An experimental test of combinatorial information markets. *Journal of Economic Behavior & Organization*, 69(2), 182–189.

Maniadis, Z., Tufano, F.,& List, J.A. (2014). One swallow doesn’t make a summer: New evidence on anchoring effects. *The American Economic Review*, 104(1), 277–90.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, 65(11), 2271–2279.

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9), 654.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Van der Laan, M. (2014). Promoting Transparency in Social Science Research. *Science* (New York, N.Y.), 343 (6166), 30–31.

Miyamoto, Y., & Kitayama, S. (2002). Cultural variation in correspondence bias: the critical role of attitude diagnosticity of socially constrained behavior. *Journal of personality and social psychology*, 83(5), 1239.

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, 26(5), 653–684.

Nuzzo, R. (2014). Statistical errors. *Nature*, 506(7487), 150–152.

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6), 657–660.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).

O'Leary, D. E. (2011). Prediction markets as a forecasting tool. Advances in business and management forecasting, 8, 169–184.

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, 7(6), 531–536.

Paul, R. J., & Weinbach, A. P. (2009). Sportsbook Behavior in the Ncaa Football Betting Market: Tests of the Traditional and Levitt Models of Sportsbook Behavior. *The Journal of Prediction Markets*, 3(2), 21–37.

Plott, C. R.,& Chen, K.-Y. (2002). Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem. Working

PaperNo. 1131, Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena.

Plott, C. R., & Sunder, S. (1982). Efficiency of Experimental Security Markets with Insider Information: An Application of Rational-Expectations Models. *Journal of Political Economy*, 90(4), 663–98.

Plott, C. R., & Sunder, S. (1988). Rational Expectations and the Aggregation of Diverse Information in Laboratory Security Markets. *Econometrica*, 56(5), 1085–1118.

Rhode, P. W. (2009). The emergence of prediction markets within business firms: a skeptical perspective from an intrigued academic. *The Journal of Prediction Markets*, 3(1), 87–88.

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of personality and social psychology*, 95(2), 293.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638-641. doi: 10.1037/0033-2909.86.3.638

Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3), 279–301.

Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science*, 12(3), 185–190.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336), 783–801.

Savani, K., Markus, H. R., Naidu, N. V. R., Kumar, S., & Berlia, N. (2010). What counts as a choice? US Americans are more likely than Indians to construe actions as choices. *Psychological Science*, 21(3), 391–398.

Schwarz, N., Strack, F., & Mai, H.-P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public opinion quarterly*, 55(1), 3–23.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & cognition*, 21(4), 546–556.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). P-Curve: A Key to the File Drawer. *Journal of Experimental Psychology*: General, Forthcoming.

Smith, V. L. (1962). An Experimental Study of Competitive Market Behavior. *Journal of Political Economy*, 70(2), 111–137.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, 54, 30-34.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on Psychological Science*, 7(6), 670–688.

Tversky, A., & Gati, I. (1978). Studies of similarity. Cognition and categorization, 1(1978), 79–98.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458.

Van Lange, P. A., De Bruin, E., Otten, W., & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence. *Journal of personality and social psychology*, 73(4), 733.

Wolfers, J., & Leigh, A. (2002). Three tools for forecasting federal elections: lessons from 2001. *Australian Journal of Political Science*, 37(2), 223–240.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, 18(2), 107–126.

Zaval, L., Keenan, E. A., Johnson, E. J., & Weber, E. U. (2014). How warm days increase belief in global warming. *Nature Climate Change*, 4(2), 143–147.

Zhong, C.-B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, 313(5792), 1451–1452.

APPENDIX

Below we describe the motivation for excluding four of the Many Labs 2 studies from our results (study 1, 7, 19, and 20), although we collected survey and prediction market data also for these four studies. In Table A1 we present the hypotheses for all the 28 Many Labs 2 studies as presented to participants in the survey and prediction markets. In Table A2 bin market beliefs about replication relative effect sizes are reported.

For four of the Many Labs 2 studies, the original studies focused on cultural differences in effect sizes between different samples (a focus of Many Labs 2 was to study to what extent effect sizes vary between samples and cultures). In study 1 the original effect went in different directions in the US and Hong Kong samples; in study 7 the original effect was significant in the US sample, but not in the Japanese sample; in study 19 the original effect was significant in the Indian sample, but not in the US sample; and in study 20 the effect was significant in the East Asian sample, but not in the sample of European Americans. In the prediction market (and survey) we therefore described the hypotheses to trade on as the results in the samples where the original studies found a significant effect for studies 7 (the US sample), study 19 (the Indian sample), and study 20 (the East Asian sample). For study 1, the hypothesis to trade on was described as the interaction testing for a significant cultural difference. However, the hypotheses for these four studies may not have been clear to the participants in the prediction markets and survey, especially as these tests were not clearly defined in the Many Labs 2 research protocol that participants were referred to for further information about the tested hypotheses. The results of these four hypotheses were also not reported in the final Many Labs 2 paper (Klein et al. 2018), with the exception of study 19. The result for the Indian subsample for study 19 was reported in the paper, but this sample size was only n=122, which is well below the sample sizes that participants were informed about in the markets and survey. The final Many Labs 2 paper also did not report the results for the US sample in study 7 or the East Asian sample in study 20 (Klein et al. 2018). In the Many Labs 2 paper there are also no results for the interaction test in study 1 (Klein et al. 2018).[25]

In the Many Labs 2 paper the authors instead reported results separately in these four studies for WEIRD (Western, Educated, Industrialized, Democratic societies) and less WEIRD samples, to capture the cultural differences observed in the original studies (Klein et al. 2018). But this classification was made after the Many Labs 2 data collection and was not available to us when we implemented the prediction markets and survey. The ambiguities surrounding these four studies make it difficult to interpret the prediction market and survey results and compare them to replication outcomes. We therefore excluded these four studies from the results.[26] Out of these four studies, the Many Labs 2 paper concluded that studies 1, 7, and 20 replicated and that study 19, did not replicate; based on these results the prediction markets predicted the results of 3 out of 4 (75%) studies correctly, which is the same

[25] The interaction test as stated in the hypothesis for this study was based on whether participants self-reported that wealthier people in their hometown live in the north or live in the south; and the hypothesis was that the tested effect would be larger for those that stated that wealthier people live in the north (like in the US sample) compared to those that stated that wealthier people live in the south (like in the Hong Kong sample). In the Many Labs 2 protocol they refer to a test of this hypothesis based on this self-reported question, but it is unclear exactly how the test will be carried out and in the published Many Labs 2 paper they only report this result for participants who self-reported that wealthier people in their hometown live in the north (and not the interaction comparing this result to those who self-reported that the wealthier people in their hometown live in the south) (Klein et al. 2018).

[26] An additional argument for excluding study 20 is that there was a mistake in the phrasing of this hypothesis in the information given to participants in the prediction markets and survey. The last part of this hypothesis read “is higher when the categorization criteria is “similar to” rather than “belongs to”. It should have read “is higher when the categorization criteria is “belongs to” rather than “similar to”.” It is thus possible that participants may have misinterpreted the direction of this hypothesis, although they may also have been informed by the Many Labs 2 research protocol and the original article.

rate as for the 24 studies reported below.[27] The survey predicted 2 out of these four correctly, which is slightly lower than the 67% rate for the 24 studies analyzed in the rest of the paper.[28]

Table A1. Directional hypotheses as presented to traders

Study	Directional Hypothesis
Study 1 Huang & Cho (2014)	When asked to indicate on a map of a fictional city where either a low-SES or a high-SES person might live there is a larger tendency to indicate more northern locations for the high-SES persons than for the low-SES person among participants who state that wealthier people in their hometown live in the North compared to participants who state that wealthier people in their hometown live in the South. Effect size (Cohen's *d*): 0.68 *P*-value: <0.001
Study 2 Kay, Laurin, Fitzsimons & Landau (2014)	Reading a description of a natural event (leaves growing on a tree) in a scenario as structured rather than random makes participants more willing to pursue their stated most important long-term goal. Effect size (Cohen's *d*): 0.50 *P*-value: 0.05
Study 3 Alter, Oppenheimer, Epley & Eyre (2007)	When presenting syllogisms in a hard to read font English speaking participants (i.e., restricted sample) solve more moderately difficult ones correctly than when presenting them in an easy-to-read font. Effect size (Cohen's *d*): 0.64 *P*-value: 0.051
Study 4 Graham, Haidt & Nosek (2009)	Compared to participants on the political left, participants on the political right deem “binding” moral concepts (that emphasize concerns for the ingroup, authority, and purity) to be more relevant for moral judgments. Effect size (Cohen's *d)*: 0.52 *P*-value: <0.001
Study 5 Rottenstreich & Hsee (2001)	When choice implementation is uncertain participants choose an affectively attractive option (kiss from favorite movie star) over a financially attractive one (money) to a higher degree than when choice implementation is certain. Effect size (Cohen's *d*): 0.75 *P*-value: 0.027
Study 6 Bauer, Wilkie, Kim & Bodenhausen (2012)	Participants in a hypothetical water conservation dilemma report less trust toward others to conserve water when they and others are referred to as “consumers” as opposed to “individuals.” Effect size (Cohen's *d*): 0.89 *P*-value: <0.001

[27] In the Many Labs 2 paper the authors did separately report the result for the Indian sample in study 19 (Klein et al. 2018), which corresponds to the way the prediction market and survey phrased this hypothesis. The study did not replicate in the Indian sample, which correspond with their overall conclusion about replication for this study in the Many Labs 2 paper (Klein et al. 2018). This was correctly predicted by the markets, but not the survey. However, this replication test was only based on n=122 Indian participants and hence it is not very informative about whether this hypothesis is supported.

[28] The prediction markets and survey (weighted survey) results for these four studies for the binary outcomes were: study 1: market beliefs=0.58, survey beliefs=0.57 (0.54); study 7: market beliefs=0.80, survey beliefs=0.76 (0.79); study 19: market beliefs=0.40, survey beliefs=0.53 (0.57); study 20: market beliefs=0.40, survey beliefs=0.46 (0.48), The prediction markets and survey (weighted survey) results for these four papers for the effect size markets were: study 1: market beliefs=0.65, survey beliefs=0.22 (0.21); study 7: market beliefs=0.69, survey beliefs=0.40 (0.42); study 19: market beliefs=0.66, survey beliefs=0.39 (0.41); study 20: market beliefs=0.64, survey beliefs=0.22 (0.23).

Study 7
Miyamoto & Kitayama (2002)

American participants (i.e. restricted sample) asked to infer a student’s true attitude toward the death penalty after having read an essay in which the student either argues for or against capital punishment (as assigned by the teacher) are influenced by the argued position (when controlling for the extent to which they thought the student’s behavior was constrained by the assignment). Participants are more likely to infer that a person’s true attitude is for capital punishment if they were assigned to write an essay for capital punishment than against it.

Effect size (Cohen's *d*): 1.78 *P*-value: <0.001

Study 8
Inbar, Pizarro, Knobe & Bloom (2009)

When a scenario describes a music video director as encouraging gay kissing in particular rather than kissing in general, the correlation between participants’ disgust sensitivity and assessments of the director’s intentionality is stronger.

Effect size (Cohen's *d*): 0.65 *P*-value: 0.034

Study 9
Critcher & Gilovich (2008)

Participants asked to predict the relative popularity of a new cell phone between two geographical regions give lower estimates of proportion sold in the home region when the cell phone is named P17 compared to when it is named P97.

Effect size (Cohen's *d*): 0.30 *P*-value: 0.035

Study 10
Van Lange, Otten, De Bruin & Joireman (1997)

There is a positive correlation between greater prosocial orientation (as measured by the SVO slider measure) and number of siblings.

P-value: <0.01

Study 11
Hauser, Cushman, Young, Kang-Xing Jin & Mikhail (2007)

Participants judge the moral permissibility of either changing the track of an out-of-control train and thus killing one person instead of five or pushing a large man with a backpack (from a bridge) in front of the train and thus killing him but saving the five. Participants (who have no prior experience with similar tasks) given the first scenario judge the action as being more morally permissible than those given the second scenario.

Effect size (Cohen's *d*): 2.51 *P*-value: <0.001

Study 12
Anderson, Kraus, Galinsky & Keltner (2012)

Participants who are made to feel high in sociometric status (or SMS, relating to interpersonal wealth) through encouragement to compare themselves to and imagine themselves interacting with people who are very low in SMS subsequently report higher subjective well-being than participants made to feel low in SMS by the same (but opposite) procedure.

Effect size (Cohen's *d*): 0.57 *P*-value: 0.003

Study 13
Ross, Greene & House (1977)

Participants’ estimates of the percentage of peers who would choose the first option in a dichotomous choice situation in a supermarket scenario are higher when the first option is also the participant’s own choice, compared to when it is not.

Effect size (Cohen's *d*): 0.79 *P*-value: <0.001

Study 14
Ross, Greene & House (1977)

Participants’ estimates of the percentage of peers who would choose the first option in a dichotomous choice situation in a traffic scenario are higher when the first option is also the participant’s own choice, compared to when it is not.

Effect size (Cohen's *d*): 0.79 *P*-value: <0.001

Study 15
Giessner & Schubert (2007)

Participants perceive a male manager as having more power when the organization chart connects the manager to his team below with a long vertical line (7 cm) instead of a short one (2 cm).

Effect size (Cohen's *d*): 0.56 *P*-value: 0.032

Study16
Tversky & Kahneman (1981)

Participants imagining themselves buying two items, one cheap ($30) and one expensive ($250), are more willing to go to a different branch of store 20 minutes away when told they could save a dollar amount ($10) on the cheap item, then when told they could save the same dollar amount on the expensive item.

Effect size (Cohen's *d*): 0.88 *P*-value: <0.001

Study 17
Hauser, Cushman, Young, Kang-Xing Jin & Mikhail (2007)

Participants judge the moral permissibility of throwing a switch to stop an out-of-control train on a set of temporary tracks in order to save five men walking on the main tracks. The train is either stopped by a man standing with his back turned or by a heavy object in front of which a man is standing with his back turned. Participants (who have no prior experience with similar tasks) given the second scenario (man in front of object) judge the action as being more morally permissible than those given the first scenario (man only).

	Effect size (Cohen's *d*): 0.34 *P*-value: <0.001
Study 18 Risen & Gilovich (2008)	Participants judge the likelihood of a negative outcome (being asked to answer a question in front of a large lecture) to be higher when imagining themselves tempting fate (by not having done the reading), compared to when not tempting fate (by having done the reading). Effect size (Cohen's *d*): 0.78 *P*-value: 0.004
Study 19 Savani, Markus, Naidu, Kumar & Berlia (2010)	Indian participants (i.e., restricted sample) who consider either interpersonal actions (such as shopping for someone else) or personal actions (such as shopping for yourself) are more likely to perceive the former as choices, when controlling for their perceived importance of the action. Effect size (Cohen's *d*): 0.33 *P*-value: <0.001
Study 20 Norenzayan, Smith, Kim & Nisbett (2002)	In a task requiring East Asian participants (i.e. restricted sample) to categorize targets into one of two groups based either on which group the target "belongs to" or on which group the target is more "similar to" the propensity to use rule-based categorization (as opposed to family resemblance) is higher when the categorization criteria is “similar to” rather than “belongs to”. Effect size (Cohen's *d*): 1.01 *P*-value: 0.050
Study 21 Hsee (1998)	Participants imagining themselves receiving either a less expensive, but high-priced item compared to other items in its category, or a more expensive, but low-priced Item compared to other items in its category, perceive the gift giver as being more generous when receiving the less expensive gift. Effect size (Cohen's *d*): 0.69 *P*-value: 0.002
Study 22 Gray & Wegner (2009)	Participants perceive a person of high moral agency (an adult man) to be more responsible for an accident (knocking over a tray of glasses) that is harmful to a person of low moral agency (a baby) compared to when the scenario is reversed. Effect size (Cohen's *d*): 0.81 *P*-value: 0.001
Study 23 Zhong & Liljenquist (2006)	Participants who retype a first-person account of an unethical story subsequently give higher desirability ratings for cleaning products (e.g., soap, toothpaste) compared to participants who retype a first-person account of an ethical story. Effect size (Cohen's *d*): 1.06 *P*-value: 0.014
Study 24 Schwarz, Strack & Mai (1991)	The correlation between participants’ ratings on two life satisfaction questions, one specific and one general, is stronger when the specific question is asked first. Effect size (Cohen's *d*): 0.48 *P*-value: <0.001
Study 25 Shafir (1993)	When deciding either to “award” or “deny” custody of a child to one of two parents, one having both more strongly positive and more strongly negative characteristics (extreme) than the other parent (average), the combined proportion of participants who award or deny custody to the extreme parent is larger than 100% (as happens when, for example, participants are more likely to both award and deny custody to the extreme parent). Effect size (Cohen's *d*): 0.42 *P*-value: 0.021
Study 26 Zaval, Keenan, Johnson & Weber (2014)	English speaking participants express a stronger belief in global warming after having performed a scrambled sentence task when the task uses sentences containing words related to heat rather than cold. Effect size (Cohen's *d*): 0.30 *P*-value: 0.039
Study 27 Knobe (2003)	Participants read vignettes about agents whose behavior have side effects that can either be helpful or harmful but are of no importance to the agent. Participants who read about harmful side effects are more likely to believe that the agent brought about the side effect intentionally compared to participants who read about helpful side effects. Effect size (Cohen's *d*): 1.46 *P*-value: <0.001
Study 28 Tversky & Gati (1978)	Participants who are presented with country pairs in which one country is more prominent than the other and are asked to rate the first country’s similarity to the second give higher similarity ratings to the pairs in which the more prominent country is listed second. For example, participants who are asked, “How similar is Mexico to the U.S.A.?”

give higher similarity ratings than participants who are asked, “How similar is the U.S.A. to Mexico?” (where the U.S.A. is more prominent).

Effect size (Cohen's *d*): 1.31 *P*-value: 0.008

Note: As explained in this appendix, studies 1, 7, 19 and 20 which focused on cultural differences were excluded from the results in our paper due to the ambiguities surrounding these four hypotheses. Note also that the Many Labs 2 did not estimate any effect size for study 10, and this study was therefore not included in the effect size markets.

Table A2. Bin market beliefs about replication relative effect sizes.

Study	Market beliefs	Bin 1 [-∞, -0.25]	Bin 2 [-0.25, 0.25]	Bin 3 [0.25, 0.75]	Bin 4 [0.75, 1.25]	Bin 5 [1.25, 1.75]	Bin 6 [1.75, +∞]
Study 2: Kay et al., (2014)	0.580	0.076	0.321	0.211	0.186	0.168	0.037
Study 3: Alter et al., (2007)	0.513	0.127	0.292	0.236	0.156	0.156	0.035
Study 4: Graham, Haidt & Nosek (2009)	0.683	0.034	0.181	0.368	0.249	0.137	0.031
Study 5: Rottenstreich & Hsee (2001)	0.634	0.086	0.229	0.271	0.196	0.178	0.040
Study 6: Bauer et al.(2012)	0.625	0.089	0.217	0.290	0.199	0.167	0.037
Study 8: Inbar et al., (2009)	0.668	0.044	0.264	0.264	0.207	0.181	0.040
Study 9: Critcher & Gilovich (2008)	0.601	0.038	0.354	0.222	0.177	0.171	0.038
Study 11: Hauser et al., (2007) st. 1	0.781	0.030	0.149	0.323	0.290	0.147	0.062
Study 12: Anderson et al., (2012)	0.666	0.056	0.213	0.311	0.221	0.163	0.036
Study 13: Ross, Greene & House (1977) st.1	0.713	0.041	0.177	0.354	0.210	0.182	0.037
Study 14: Ross, Greene & House (1977) st. 2	0.687	0.038	0.205	0.348	0.202	0.169	0.038
Study 15: Giessner & Schubert (2007)	0.575	0.081	0.300	0.243	0.176	0.163	0.036
Study 16: Tversky & Kahneman (1981)	0.733	0.034	0.186	0.302	0.279	0.158	0.042
Study 17: Hauser et al., (2007) st. 2	0.737	0.035	0.183	0.289	0.296	0.163	0.035
Study 18: Risen & Gilovich (2008)	0.723	0.041	0.205	0.287	0.240	0.185	0.041
Study 21: Hsee (1998)	0.699	0.040	0.224	0.290	0.229	0.178	0.040
Study 22: Gray & Wegner (2009)	0.721	0.035	0.173	0.334	0.264	0.158	0.035
Study 23: Zhong & Liljenquist (2006)	0.503	0.121	0.327	0.203	0.160	0.157	0.034
Study 24: Schwarz, Strack & Mai (1991)	0.699	0.038	0.187	0.360	0.207	0.170	0.038
Study 25: Shafir (1993)	0.716	0.044	0.251	0.218	0.247	0.196	0.044
Study 26: Zaval et al., (2014)	0.588	0.036	0.351	0.247	0.167	0.162	0.036
Study 27: Knobe (2003)	0.799	0.042	0.196	0.215	0.257	0.250	0.040
Study 28: Tversky & Gati (1978)	0.790	0.051	0.198	0.230	0.229	0.227	0.066

Note: Study 10 was not part of the effect size markets, and studies 1, 7, 19, and 20 were excluded, as explained in this appendix.

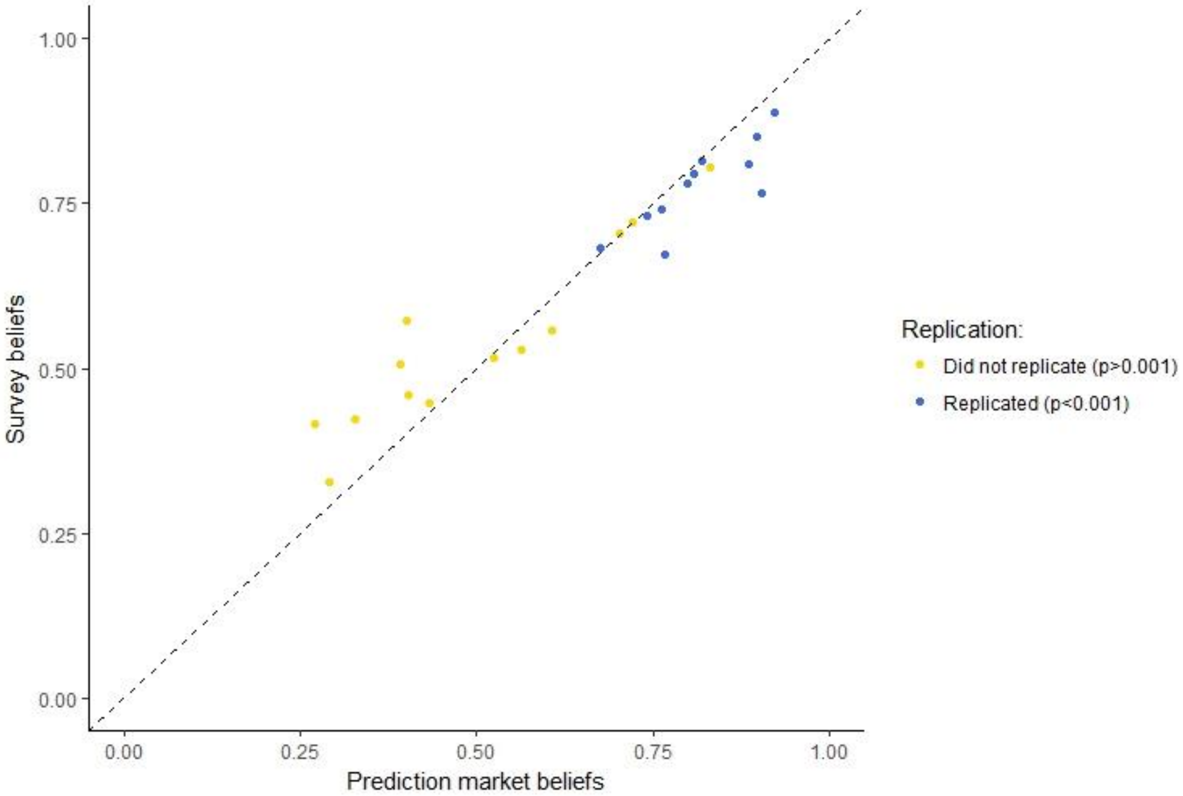

FIGURE A1: Prediction market and survey beliefs for the replication probability.

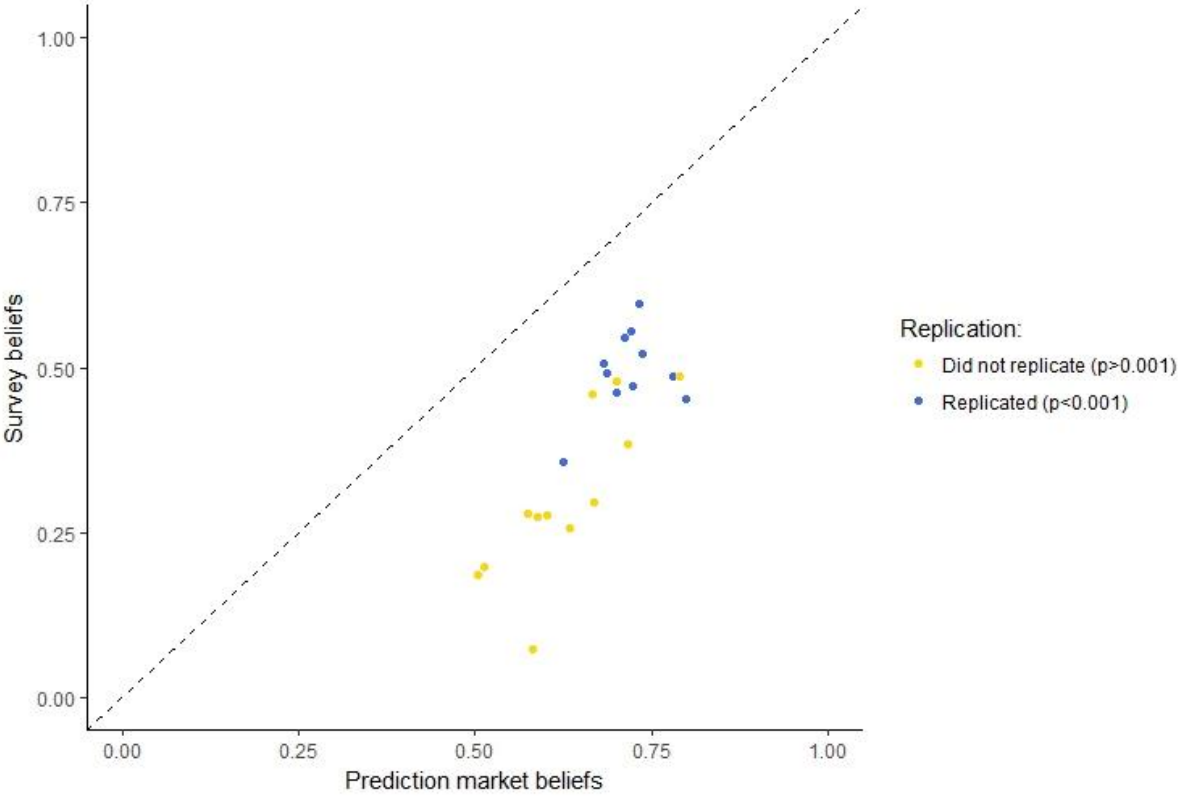

FIGURE A2: Prediction market and survey beliefs for the relative effect size.