

Using prediction markets to estimate the reproducibility of scientific research

Anna Dreber^{a,1,2}, Thomas Pfeiffer^{b,c,1}, Johan Almenberg^d, Siri Isaksson^a, Brad Wilson^e, Yiling Chen^f, Brian A. Nosek^{g,h}, and Magnus Johannesson^a

^aDepartment of Economics, Stockholm School of Economics, SE-113 83 Stockholm, Sweden; ^bNew Zealand Institute for Advanced Study, Massey University, Auckland 0745, New Zealand; ^cWissenschaftskolleg zu Berlin–Institute for Advanced Study, D-14193 Berlin, Germany; ^dSveriges Riksbank, SE-103 37 Stockholm, Sweden; ^eConsensus Point, Nashville, TN 37203; ^fJohn A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; ^gDepartment of Psychology, University of Virginia, Charlottesville, VA 22904; and ^hCenter for Open Science, Charlottesville, VA 22903

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved October 6, 2015 (received for review August 17, 2015)

Concerns about a lack of reproducibility of statistically significant results have recently been raised in many fields, and it has been argued that this lack comes at substantial economic costs. We here report the results from prediction markets set up to quantify the reproducibility of 44 studies published in prominent psychology journals and replicated in the Reproducibility Project: Psychology. The prediction markets predict the outcomes of the replications well and outperform a survey of market participants' individual forecasts. This shows that prediction markets are a promising tool for assessing the reproducibility of published scientific results. The prediction markets also allow us to estimate probabilities for the hypotheses being true at different testing stages, which provides valuable information regarding the temporal dynamics of scientific discovery. We find that the hypotheses being tested in psychology typically have low prior probabilities of being true (median, 9%) and that a "statistically significant" finding needs to be confirmed in a well-powered replication to have a high probability of being true. We argue that prediction markets could be used to obtain speedy information about reproducibility at low cost and could potentially even be used to determine which studies to replicate to optimally allocate limited resources into replications.

reproducibility | replications | prediction markets

The process of scientific discovery centers on empirical testing of research hypotheses. A standard tool to interpret results in statistical hypothesis testing is the P value. A result associated with a P value below a predefined significance level (typically 0.05) is considered "statistically significant" and interpreted as evidence in favor of a hypothesis. However, concerns about the reproducibility of statistically significant results have recently been raised in many fields including medicine (1–3), neuroscience (4), genetics (5, 6), psychology (7–11), and economics (12, 13). For example, an industrial laboratory could only reproduce 6 out of 53 key findings from "landmark" studies in preclinical oncology (2) and it has been argued that the costs associated with irreproducible preclinical research alone are about US\$28 billion a year in the United States (3). The mismatch between the interpretation of statistically significant findings and a lack of reproducibility threatens to undermine the validity of statistical hypothesis testing as it is currently practiced in many research fields (14).

The problem with inference based on P values is that a P value provides only partial information about the probability of a tested hypothesis being true (14, 15). This probability also depends on the statistical power to detect a true positive effect and the prior probability that the hypothesis is true (14). Lower statistical power increases the probability that a statistically significant effect is a false positive (4, 14). Statistically significant results from small studies are therefore more likely to be false positives than statistically significant results from large studies. A lower prior probability for a hypothesis to be true similarly increases the probability that a statistically significant effect is a false positive

(14). This problem is exacerbated by publication bias in favor of speculative findings and against null results (4, 16–19).

Apart from rigorous replication of published studies, which is often perceived as unattractive and therefore rarely done, there are no formal mechanisms to identify irreproducible findings. Thus, it is typically left to the judgment of individual researchers to assess the credibility of published results. Prediction markets are a promising tool to fill this gap, because they can aggregate private information on reproducibility, and can generate and disseminate a consensus among market participants. Although prediction markets have been argued to be a potentially important tool for assessing scientific hypotheses (20–22)—most notably in Robin Hanson's paper "Could Gambling Save Science? Encouraging an Honest Consensus" (20)—relatively little has been done to develop potential applications (21). Meanwhile, the potential of prediction markets has been demonstrated in a number of other domains, such as sports, entertainment, and politics (23–26).

We tested the potential of using prediction markets to estimate reproducibility in conjunction with the Reproducibility Project: Psychology (RPP) (9, 10). The RPP systematically replicated studies from a sampling frame of three top journals in psychology. To investigate the performance of prediction markets in this context, a first set of prediction markets were implemented in November 2012 and included 23 replication studies scheduled to be completed in the subsequent 2 mo, and a second set of

Significance

There is increasing concern about the reproducibility of scientific research. For example, the costs associated with irreproducible preclinical research alone have recently been estimated at US\$28 billion a year in the United States. However, there are currently no mechanisms in place to quickly identify findings that are unlikely to replicate. We show that prediction markets are well suited to bridge this gap. Prediction markets set up to estimate the reproducibility of 44 studies published in prominent psychology journals and replicated in The Reproducibility Project: Psychology predict the outcomes of the replications well and outperform a survey of individual forecasts.

Author contributions: A.D., T.P., J.A., B.A.N., and M.J. designed research; A.D., T.P., J.A., S.I., B.W., Y.C., B.A.N., and M.J. performed research; A.D., T.P., J.A., and M.J. analyzed data; A.D., T.P., J.A., and M.J. wrote the paper.

Conflict of interest statement: Consensus Point employs B.W. and provided the online market interface used in the experiment. The market interface is commercial software.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Open Science Framework database, <https://osf.io/jymht>.

See Commentary on page 15267.

¹A.D. and T.P. contributed equally to this work.

²To whom correspondence should be addressed. Email: anna.dreber@hhs.se.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1516179112/-DCSupplemental.

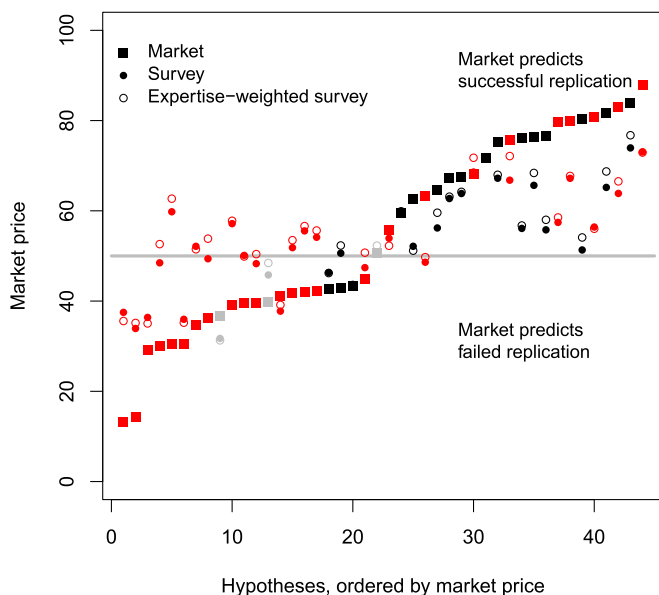


Fig. 1. Prediction market performance. Final market prices and survey predictions are shown for the replication of 44 publications from three top psychology journals. The prediction market predicts 29 out of 41 replications correctly, yielding better predictions than a survey carried out before the trading started. Successful replications (16 of 41 replications) are shown in black, and failed replications (25 of 41) are shown in red. Gray symbols are replications that remained unfinished (3 of 44).

prediction markets were implemented in October 2014 and included 21 replication studies scheduled to be completed before the end of December 2014. The prediction markets were active for 2 wk at each of these occasions.

For each of the replication studies, participants could bet on whether or not the key original result would be replicated. Our criterion for a successful replication was a replication result, with a P value of less than 0.05, in the same direction as the original result. In one of the studies, the original result was a negative finding, and successful replication was thus defined as obtaining a negative (i.e., statistically nonsignificant) result in the replication. Information on the original study and the setup of the replication were accessible to all participants.

In the prediction markets, participants traded contracts that pay \$1 if the study is replicated and \$0 otherwise. This type of contract allows the price to be interpreted as the predicted probability of the outcome occurring. This interpretation of the price is not without caveats (27) but has an advantage of being simple and reasonably robust (28), especially in settings where traders' initial endowments are the same and traders' bets are relatively small. Invitations to participate in the prediction markets were sent to the email list of the Open Science Framework, and for the second set of markets also to the email list of the RPP collaboration. Participants were not allowed to bet in those markets where they were involved in carrying out the replication. In the first set of prediction markets, 49 individuals signed up and 47 of these actively participated; in the second set, 52 individuals signed up and 45 of these actively participated. Before the markets started, participants were asked in a survey for their subjective probability of each study being replicated. Each participant was endowed with US\$100 for trading.

Results

The prediction markets functioned well in an operational sense. Participation was broad, i.e., trading was not dominated by a small subset of traders or concentrated to just a few of the markets. In total, 2,496 transactions were carried out. The number of

transactions per market ranged from 28 to 108 (mean, 56.7), and the number of active traders per market ranged from 18 to 40 (mean, 26.7). We did not detect any market bias regarding bets on success ("long positions") or failure ("short positions") to replicate the original results. In the final portfolios held at market closing time (*Supporting Information*), we observed approximately the same number of bets on success and failure.

The mean prediction market final price is 55% (range, 13–88%), implying that about half of the 44 studies were expected to replicate. Out of the 44 scientific studies included in the prediction markets, the replications were completed for 41 of the studies, with the remaining replications being delayed. Of the 41 completed, 16 studies (39%) replicated and 25 studies (61%) did not replicate according to the market criterion for a successful replication (*Supporting Information*).

We evaluate the performance of the markets in three ways. We test whether the market prices are informative; if the market prices can be interpreted as probabilities of replication; and if the prediction markets predict the replication outcomes better than a survey measure of beliefs. When interpreting a market price larger than 50% as predicting successful replication and a market price smaller than 50% as predicting failed replication, informative markets are expected to correctly predict more than 50% of the replications. We find that the prediction markets correctly predict the outcome of 71% of the replications (29 of 41 studies; Fig. 1), which is significantly higher than 50% (one-sample binomial test; $P = 0.012$).

Interpreting the prediction market prices as probabilities means that not all markets with a price larger (smaller) than 50% are expected to correspond to successful (failed) replications. The expected prediction rate of the markets depends on the distribution of final market prices, which in our study implies that 69% of the outcomes are expected to be predicted correctly. This is very close to the observed value of 71%. To formally test whether prediction market prices can be interpreted as probabilities of replication, we estimated a linear probability model (with robust SEs) with the outcome of the replication as a function of the prediction market price. If market prices equal replication probabilities, the coefficient of the market price variable should be equal to 1 and the constant in the regression should be equal to zero. The coefficient of the market price variable is 0.995, which is significantly different from zero ($P = 0.003$), but not significantly different from 1 ($P = 0.987$). The constant (-0.167) is not significantly different from zero ($t = -1.11$, $P = 0.276$).

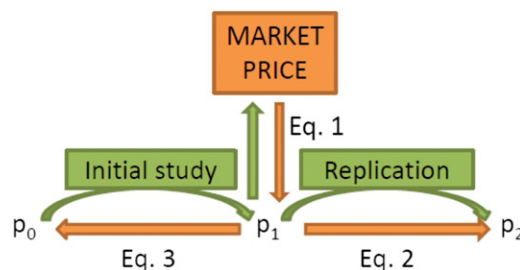


Fig. 2. Relationship between market price and prior and posterior probabilities p_0 , p_1 , and p_2 of the hypothesis under investigation. Bayesian inference (green arrows) assigns an initial (prior) probability p_0 to a hypothesis, indicating its plausibility in absence of a direct test. Results from an initial study allows this prior probability to be updated to posterior p_1 , which in turn determines the chances for the initial result to hold up in a replication, and thus the market price in the prediction market. Once the replication has been performed, the result can be used to generate posterior p_2 . Observing the market price, and using the statistical characteristics of the initial study and the replication, we can thus reconstruct probabilities p_1 , p_2 , and p_0 . Detailed calculations are presented in *Supporting Information*.

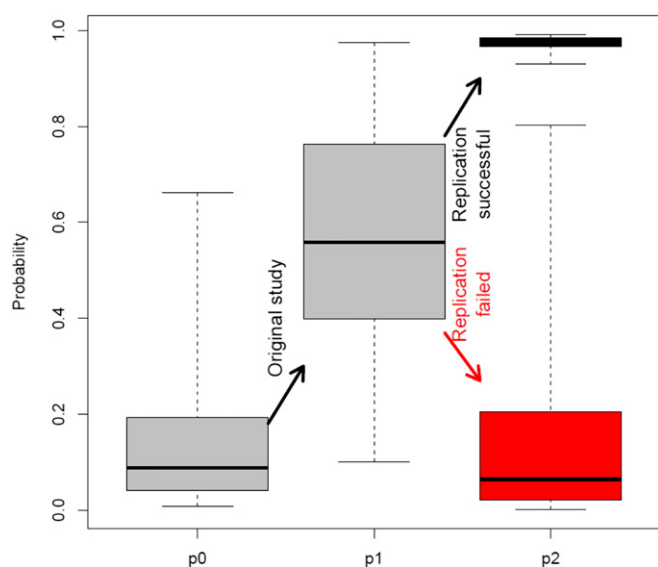


Fig. 3. Probability of a hypothesis being true at three different stages of testing: before the initial study (p_0), after the initial study but before the replication (p_1), and after replication (p_2). “Error bars” (or whiskers) represent range, boxes are first to third quartiles, and thick lines are medians. Initially, priors of the tested hypothesis are relatively low, with a median of 8.8% (range, 0.7–66%). A positive result in an initial publication then moves the prior into a broad range of intermediate levels, with a median of 56% (range, 10–97%). If replicated successfully, the probability moves further up, with a median of 98% (range, 93.0–99.2%). If the replication fails, the probability moves back to a range close to the initial prior, with a median of 6.3% (range, 0.01–80%).

The prediction market can also be compared with the pretrading survey of participants’ beliefs about the probability of replication. A simple average of the survey correctly predicts 58% of outcomes (23 of 40; Fig. 1; survey data are missing for one market), which is not significantly different from 50% (one-sample binomial test; $P = 0.429$). A weighted average, using self-reported expertise as weights, correctly predicts 50% (20 of 40) of outcomes, which is not significantly different from 50% (one-sample binomial test; $P = 1.00$). The absolute prediction error is significantly lower for the prediction market than for both the pretrading survey (paired t test, $n = 40$, $t = -2.558$, $P = 0.015$) and the weighted survey (paired t test, $n = 40$, $t = -2.727$, $P = 0.010$; see [Supporting Information](#) for a more detailed comparison of the prediction market and survey responses). The prediction market thus outperforms the survey measure of beliefs.

The above results suggest that the prediction markets generate good estimates of the probability that a published result will be replicated. Note that the probability of successful replication is not the same thing as the probability of a tested hypothesis being true. The probability of a tested hypothesis being true, also referred to as the positive predictive value or PPV (4), can however be estimated from the market price (Fig. 2). Using information about the power and significance levels of the original study and the replications (see [Supporting Information](#) for details), it can be estimated for three stages of the testing process: the prior probability (p_0) before observing the outcome of the initial study; the probability after observing the result of the initially published study (p_1); and the probability after observing the outcome of the replication (p_2). A summary of the results of these estimations are shown in Fig. 3; a more detailed breakdown is given in [Supporting Information](#).

Our analysis reveals priors (p_0) for the 44 studies ranging from 0.7% to 66% with a median (mean) of 8.8% (13%). This relatively low average prior may reflect that top psychology journals

focus on publishing surprising findings, i.e., positive findings on relatively unlikely hypotheses. The probability that the research hypothesis is true after observing the positive finding in the first study (p_1) ranges from 10% to 97% with a median (mean) of 56% (57%) for the 44 studies. This estimate implies that about 43% of statistically significant research findings published in these top psychology journals can be expected to be false positives.

For the 41 studies replicated so far, we can also estimate the posterior probability that the research finding is true contingent on observing the result of the replication (p_2). This probability ranges between 93.0% and 99.2% with a median (mean) of 98% (97%) for the 16 studies whose result was replicated, and between 0.1% and 80% with a median (mean) of 6.3% (15%) for the 25 studies that were not replicated.

These results show that prediction markets can give valuable insights into the dynamics of information accumulation in a research field. Eliciting priors in this manner allows us to evaluate whether hypotheses are tested appropriately in a given research field. A common, but incorrect, interpretation of a published result with a $P < 0.05$ is that it implies a 95% probability of the research hypothesis being true. Interestingly, our findings imply that to achieve such a high probability of the research hypothesis being true, a “statistically significant” positive finding needs to be confirmed in a well-powered replication. This illustrates the importance of replicating positive research findings before they are given high credibility. It remains to be studied how psychology compares in this aspect to other fields.

Discussion

The RPP project recently found that more than one-half of 100 original findings published in top psychology journals failed to replicate (10). Our prediction market results suggest that this relatively low rate of reproducibility should not come as a surprise to the profession, as it is consistent with the beliefs held by psychologists participating in our prediction market.

As can be seen in Fig. 1, original findings for which the market prices indicated a low probability of replication were indeed typically not replicated. However, there were also some findings that failed to replicate despite high market prices that indicated that participants had less doubts about those findings. An interesting hypothesis is that in some of these cases it was the replication itself, rather than the original finding, that failed. It would thus be particularly interesting to carry out additional replications of these studies.

Although our results suggest that prediction markets can be used to obtain accurate forecasts regarding the outcome of replications, one limitation of the approach we used in this study lays in the necessity to run replications so that there is an outcome to trade on. Some studies such as large field experiments may be very costly to replicate (29). One way to mitigate this would be to run prediction markets on a number of studies, from which a subset is randomly selected for replication after the market closes (20). Such an approach could provide quick information about reproducibility at low cost. Moreover, prediction markets could potentially be used as so-called “decision markets” (30, 31) to prioritize replication of some studies, such as those with the lowest likelihood of replication. This would generate salient and informative signals about reproducibility, and help optimizing the allocation of resources into replication.

Materials and Methods

The RPP by the Open Science Collaboration (10) sampled papers in the 2008 issues of three top psychology journals: *Journal of Personality and Social Psychology*, *Psychological Science*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. In the case of several studies in one paper, typically the last study of each paper was selected for replication.

We chose 23 studies for the first set of prediction markets and 21 studies for the second set of prediction markets, where the chosen studies were

scheduled to be replicated within 2 mo after the completion of the prediction market. For each replication, the hypothesis of the original study was summarized by one of the authors of this paper and submitted to the replication team for comments and final approval. In 1 of the 23 studies in the first prediction market, the chosen experiment was changed by the replicating researcher after the survey had been performed but before the trading started (SI ref. 34 in [Supporting Information](#)); we thus lack survey data for this study. One of the 21 studies in the second prediction market was later changed for a different experiment to be replicated (SI ref. 59 in [Supporting Information](#)), but for completeness we still include the prediction market and survey data for this study (although there are no current plans to replicate this study).

Participants in the prediction market were researchers in various fields of psychology, ranging from graduate students to professors. Fourteen participants were directly involved in one or several replication studies (15 studies in total) and were not allowed to make trades on the outcomes of these specific studies. Sixteen participants participated in both sets of prediction markets. Before the prediction market, the participants filled out a survey. For each study, participants were asked two questions. One was meant to capture their beliefs of reproducibility: "How likely do you think it is that this hypothesis will be replicated (on a scale from 0% to 100%)?" Participants were also asked about their expertise in the area: "How well do you know this topic? (not at all, slightly, moderately, very well, extremely well)." We transformed this latter measure into a 1–5 scale, and it was used to construct the weighted average belief measure from the survey.

Trading in the prediction market took place through a web-based market interface in collaboration with Consensus Point (www.consensuspoint.com/), a leading provider of prediction market research technology. Before starting to trade, participants received information about the trading procedure as well as logins. Trading accounts were initially endowed with \$100 (expressed as 10,000 "points"). These points were used to make predictions of successful replication. Predictions were made by buying and selling stocks on the hypotheses on an interface that highlighted the forecasting functionality of the market ([Supporting Information](#)). In the prediction market, participants traded contracts that pay \$1 (i.e., 100 points) if the study is replicated and \$0 otherwise. This type of contract allows the price to be interpreted as the predicted probability of the outcome occurring. For each hypothesis, participants could see the current market prediction for the probability of successful replication.

The trading platform used an automated market maker implementing a logarithmic market scoring rule (32). This algorithm offers a buying price and

a selling price at all times, ensuring that there is always a counterpart with which to trade. More specifically, the algorithm uses the net sales (s) the market maker has done so far in a market to determine the prices for a (infinitesimally small) trade as $P = \exp(s/b)/(\exp(s/b) + 1)$. To buy stocks, participants chose YES on the trading interface and entered how many points they would like to invest. For each additional point invested in a YES position, the price (and the predicted probability for successful replication) increased. To sell stocks, participants chose NO on the trading interface and entered how many points they would like to invest. For each additional point invested in a NO position, the price decreased. Participants could also buy (sell) shares by increasing (decreasing) an existing YES position, or decreasing (increasing) an existing NO position. The market maker ensures that the value of a YES share is \$1 minus the value of a NO share. Parameter b determines the liquidity and the maximal subsidies provided by the market maker and controls how strongly the market price is affected by a trade. We set the liquidity parameter to $b = 100$ (points). This means that, by investing 1,000 points (i.e., 1/10 of the initial endowment), traders can move the price of a single market from 50% to about 55%; and investing the entire initial endowment into a single market moves the price from 50% to 82%.

For the first set of prediction markets, investments were settled 5 mo after the market had closed according to actual results of the replications in the cases where the outcome was available and to market value in the cases where the replications were not yet finished. At the time of the close of the market, only eight results were known by the replicating researcher, where all replicating researchers had agreed to not share the results with anyone until after the market closed. For the second set of prediction markets, investments were similarly settled 4.5 mo after the markets had closed. At the time of the close of the second market, one result was known by the replicating researcher; all replicating researchers agreed here too to not share their results with anyone until the market had closed.

ACKNOWLEDGMENTS. We thank Agneta Berge for research assistance; and Juergen Huber, Willemien Kets, and Pranjal Mehta for comments on a previous version of the manuscript. We thank the Jan Wallander and Tom Hedelius Foundation (P2012-0002:1, P2013-0156:1, and P2015-0001:1), the Knut and Alice Wallenberg Foundation [Wallenberg Academy Fellows Grant (to A.D.)], the Swedish Foundation for Humanities and Social Sciences (NHS 14-1719:1), and the National Science Foundation (Grant CCF-0953516) for financial support.

- Prinz F, Schlange T, Asadullah K (2011) Believe it or not: How much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 10(9):712.
- Begley CG, Ellis LM (2012) Drug development: Raise standards for preclinical cancer research. *Nature* 483(7391):531–533.
- Freedman LP, Cockburn IM, Simcoe TS (2015) The economics of reproducibility in preclinical research. *PLoS Biol* 13(6):e1002165.
- Button KS, et al. (2013) Power failure: Why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14(5):365–376.
- Hewitt JK (2012) Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behav Genet* 42(1):1–2.
- Lawrence MS, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457):214–218.
- Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22(11):1359–1366.
- Carpenter S (2012) Psychology research. Psychology's bold initiative. *Science* 335(6076):1558–1561.
- Open Science Collaboration (2012) An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect Psychol Sci* 7(6):657–660.
- Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716.
- Bohannon J (2014) Psychology. Replication effort provokes praise—and "bullying" charges. *Science* 344(6186):788–789.
- Ioannidis J, Doucouliagos CJ (2013) What's to know about the credibility of empirical economics? *J Econ Surv* 27(5):997–1004.
- Maniatis Z, Tufano F, List JA (2014) One swallow doesn't make a summer: How economists (mis-) use experimental methods and their results. *Am Econ Rev* 104(1):277–290.
- Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2(8):e124.
- Nuzzo R (2014) Scientific method: Statistical errors. *Nature* 506(7487):150–152.
- Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR (1991) Publication bias in clinical research. *Lancet* 337(8746):867–872.
- Stern JM, Simes RJ (1997) Publication bias: Evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 315(7109):640–645.
- Ioannidis JPA, Munafò MR, Fusar-Poli P, Nosek BA, David SP (2014) Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends Cogn Sci* 18(5):235–241.
- Miguel E, et al. (2014) Social science. Promoting transparency in social science research. *Science* 343(6166):30–31.
- Hanson RD (1995) Could gambling save science? Encouraging an honest consensus. *Soc Epistemology* 9(1):3–33.
- Almenberg J, Kittlitz K, Pfeiffer T (2009) An experiment on prediction markets in science. *PLoS One* 4(12):e8500.
- Park I-U, Peacey MW, Munafò MR (2014) Modelling the effects of subjective and objective decision making in scientific peer review. *Nature* 506(7486):93–96.
- Wolffers J, Zitzewitz E (2004) Prediction markets. *J Econ Perspect* 18(2):107–126.
- Tziralis G, Tsiatopoulos I (2007) Prediction markets: An extended literature review. *J Pred Markets* 1(1):75–91.
- Arrow KJ, et al. (2008) Economics. The promise of prediction markets. *Science* 320(5878):877–878.
- Berg J, Forsythe R, Nelson F, Rietz T (2008) Results from a dozen years of election futures markets research. *Handbook of Experimental Economics Results*, eds Plott CR, Smith VL (North-Holland, Amsterdam), pp 742–751.
- Manski CF (2006) Interpreting the predictions of prediction markets. *Econ Lett* 91(3):425–429.
- Wolffers J, Zitzewitz E (2006) *Interpreting Prediction Market Prices as Probabilities* (National Bureau of Economic Research, Cambridge, MA), NBER Working Paper No. 12200.
- Manning WG, et al. (1987) Health insurance and the demand for medical care: Evidence from a randomized experiment. *Am Econ Rev* 77(3):251–277.
- Hanson R (2006) Decision markets for policy advice. *Promoting the General Welfare: American Democracy and the Political Economy of Government Performance*, eds Patashnik EM, Gerber AS (Brookings Institution Press, Washington, DC), pp 151–173.
- Chen Y, Kash IA, Ruberry M, Shnayder V (2014) Eliciting predictions and recommendations for decision making. *ACM Trans Econ Comput* 2(2):6:1–6:27.
- Hanson R (2007) Logarithmic market scoring rules for modular combinatorial information aggregation. *J Pred Markets* 1(1):3–15.
- Richeson JA, Trawalter S (2008) The threat of appearing prejudiced and race-based attentional biases. *Psychol Sci* 19(2):98–102.

34. Reynolds M, Besner D (2008) Contextual effects on reading aloud: Evidence for pathway control. *J Exp Psychol Learn Mem Cogn* 34(1):50–64.
35. Rule NO, Ambady N (2008) The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychol Sci* 19(2):109–111.
36. Morris AL, Still ML (2008) Now you see it, now you don't: Repetition blindness for nonwords. *J Exp Psychol Learn Mem Cogn* 34(1):146–166.
37. Shnabel N, Nadler A (2008) A needs-based model of reconciliation: Satisfying the differential emotional needs of victim and perpetrator as a key to promoting reconciliation. *J Pers Soc Psychol* 94(1):116–132.
38. Correll J (2008) 1/f noise and effort on implicit measures of bias. *J Pers Soc Psychol* 94(1):48–59.
39. Fischer P, Greitemeyer T, Frey D (2008) Self-regulation and selective exposure: The impact of depleted self-regulation resources on confirmatory information processing. *J Pers Soc Psychol* 94(3):382–395.
40. Alter AL, Oppenheimer DM (2008) Effects of fluency on psychological distance and mental construal (or why New York is a large city, but New York is a civilized jungle). *Psychol Sci* 19(2):161–167.
41. Mirman D, Magnuson JS (2008) Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *J Exp Psychol Learn Mem Cogn* 34(1):65–79.
42. Estes Z, Verges M, Barsalou LW (2008) Head up, foot down: Object words orient attention to the objects' typical location. *Psychol Sci* 19(2):93–97.
43. Nairne JS, Pandeirada JNS, Thompson SR (2008) Adaptive memory: The comparative value of survival processing. *Psychol Sci* 19(2):176–180.
44. White PA (2008) Accounting for occurrences: A new view of the use of contingency information in causal judgment. *J Exp Psychol Learn Mem Cogn* 34(1):204–218.
45. Pacton S, Perruchet P (2008) An attention-based associative account of adjacent and nonadjacent dependency learning. *J Exp Psychol Learn Mem Cogn* 34(1):80–96.
46. Pleskac TJ (2008) Decision making and learning while taking sequential risks. *J Exp Psychol Learn Mem Cogn* 34(1):167–185.
47. Masicampo EJ, Baumeister RF (2008) Toward a physiology of dual-process reasoning and judgment: Lemonade, willpower, and expensive rule-based analysis. *Psychol Sci* 19(3):255–260.
48. Janssen N, Schirm W, Mahon BZ, Caramazza A (2008) Semantic interference in a delayed naming task: Evidence for the response exclusion hypothesis. *J Exp Psychol Learn Mem Cogn* 34(1):249–256.
49. Henderson MD, de Liver Y, Gollwitzer PM (2008) The effects of an implemental mind-set on attitude strength. *J Pers Soc Psychol* 94(3):396–411.
50. Turk-Browne NB, Isola PJ, Scholl BJ, Treat TA (2008) Multidimensional visual statistical learning. *J Exp Psychol Learn Mem Cogn* 34(2):399–407.
51. Vul E, Nieuwenstein M, Kanwisher N (2008) Temporal selection is suppressed, delayed, and diffused during the attentional blink. *Psychol Sci* 19(1):55–61.
52. Vohs KD, Schooler JW (2008) The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychol Sci* 19(1):49–54.
53. Fischer P, Schulz-Hardt S, Frey D (2008) Selective exposure and information quantity: How different information quantities moderate decision makers' preference for consistent and inconsistent information. *J Pers Soc Psychol* 94(2):231–244.
54. Bressan P, Stranieri D (2008) The best men are (not always) already taken: Female preference for single versus attached males depends on conception risk. *Psychol Sci* 19(2):145–151.
55. Sahakyan L, Delaney PF, Waldum ER (2008) Intentional forgetting is easier after two "shots" than one. *J Exp Psychol Learn Mem Cogn* 34(2):408–414.
56. Lobue V, DeLoache JS (2008) Detecting the snake in the grass: Attention to fear-relevant stimuli by adults and young children. *Psychol Sci* 19(3):284–289.
57. Nurmsoo E, Bloom P (2008) Preschoolers' perspective taking in word learning: Do they blindly follow eye gaze? *Psychol Sci* 19(3):211–215.
58. Wolf ST, Insko CA, Kirchner JL, Wildschut T (2008) Interindividual-intergroup discontinuity in the domain of correspondent outcomes: The roles of relativistic concern, perceived categorization, and the doctrine of mutual assured destruction. *J Pers Soc Psychol* 94(3):479–494.
59. Förster J, Liberman N, Kuschel S (2008) The effect of global versus local processing styles on assimilation versus contrast in social judgment. *J Pers Soc Psychol* 94(4):579–599.
60. Lau GP, Kay AC, Spencer SJ (2008) Loving those who justify inequality: The effects of system threat on attraction to women who embody benevolent sexist ideals. *Psychol Sci* 19(1):20–21.
61. Stanovich KE, West RF (2008) On the relative independence of thinking biases and cognitive ability. *J Pers Soc Psychol* 94(4):672–695.
62. Lemay EP, Jr, Clark MS (2008) "Walking on eggshells": How expressing relationship insecurities perpetuates them. *J Pers Soc Psychol* 95(2):420–441.
63. Stinson DA, et al. (2008) The cost of lower self-esteem: Testing a self- and social-bonds model of health. *J Pers Soc Psychol* 94(3):412–428.
64. McCreary SM (2008) Self-handicapping, excuse making, and counterfactual thinking: Consequences for self-esteem and future motivation. *J Pers Soc Psychol* 95(2):274–292.
65. Halevy N, Bornstein G, Sagiv L (2008) "In-group love" and "out-group hate" as motives for individual participation in intergroup conflict: A new game paradigm. *Psychol Sci* 19(4):405–411.
66. Yap MJ, Balota DA, Tse CS, Besner D (2008) On the additive effects of stimulus quality and word frequency in lexical decision: Evidence for opposing interactive influences revealed by RT distributional analyses. *J Exp Psychol Learn Mem Cogn* 34(3):495–513.
67. Koo M, Fishbach A (2008) Dynamics of self-regulation: How (un)accomplished goal actions affect motivation. *J Pers Soc Psychol* 94(2):183–195.
68. Schmidt JR, Besner D (2008) The Stroop effect: Why proportion congruent has nothing to do with congruency and everything to do with contingency. *J Exp Psychol Learn Mem Cogn* 34(3):514–523.
69. Dodson CS, Darragh J, Williams A (2008) Stereotypes and retrieval-provoked illusory source recollections. *J Exp Psychol Learn Mem Cogn* 34(3):460–477.
70. Fiedler K (2008) The ultimate sampling dilemma in experience-based decision making. *J Exp Psychol Learn Mem Cogn* 34(1):186–203.
71. Berry CJ, Shanks DR, Henson RNA (2008) A single-system account of the relationship between priming, recognition, and fluency. *J Exp Psychol Learn Mem Cogn* 34(1):97–111.
72. Tamir M, Mitchell C, Gross JJ (2008) Hedonic and instrumental motives in anger regulation. *Psychol Sci* 19(4):324–328.
73. Liefvooghe B, Barrouillet P, Vandierendonck A, Camos V (2008) Working memory costs of task switching. *J Exp Psychol Learn Mem Cogn* 34(3):478–494.
74. Farris C, Treat TA, Viken RJ, McFall RM (2008) Perceptual mechanisms that characterize gender differences in decoding women's sexual intent. *Psychol Sci* 19(4):348–354.
75. Purdie-Vaughns V, Steele CM, Davies PG, Dittmann R, Crosby JR (2008) Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *J Pers Soc Psychol* 94(4):615–630.
76. Bassok M, Pedigo SF, Oskarsson AT (2008) Priming addition facts with semantic relations. *J Exp Psychol Learn Mem Cogn* 34(2):343–352.

Supporting Information

Dreber et al. 10.1073/pnas.1516179112

Here, we provide further details on the market performance; the comparison of the prediction market and survey responses; the reconstruction of the prior and posterior probabilities (p_0 , p_1 , and p_2) from the market price; the association between the market price and the statistical power; and results and data for the individual studies.

Market Performance

The overall trading volume in the first set of prediction markets ranged from 169 to 2,564 (mean, 921; median, 797) in terms of traded shares, and from 9,671 to 146,472 (mean, 51,486; median, 46,415) in terms of cash. In the second set of markets, volumes ranged from 365 to 1,155 (mean, 555; median, 506) in terms of traded shares, and from 18,721 to 67,033 (mean, 30,147; median, 27,987) in terms of cash.

We distinguish between four types of transactions: increasing a long position, reducing a long position, increasing a short position, and reducing a short position. In the first set of markets, 618 transactions were carried out to increase a long position (average volume, 12.4; median volume, 6.8), 157 to reduce a long position (average volume, 22.4; median volume, 9.8), 549 to increase a short position (average volume, 12.8; median volume, 8.8), and 156 to reduce a short position (average volume, 18.9; median volume, 9.9). In the second set of markets, 408 transactions were carried out to increase a long position (average volume, 13.8; median volume, 10.4), 77 to reduce a long position (average volume, 11.3; median volume, 5.5), 454 to increase a short position (average volume, 9.8; median volume, 6.4), and 77 to reduce a short position (average volume, 8.8; median volume, 5.9). Thus, transactions to reduce existing positions were larger in volume than transactions to enter new positions or increase existing ones; and trading into long positions and short positions showed similar patterns.

Comparison of the Prediction Market and Survey Responses

There is considerable overlap between the prediction market and survey responses (Fig. 1 and Fig. S3), suggesting that the information given in the survey is also reflected in the market. The market generated predictions over a wider range of 13–88% compared with the survey range of 32–74%; i.e., the prediction market was more informative than the survey, in the narrow sense that the survey generated predictions closer to a diffuse (noninformative) prior. This constitutes additional support for the interpretation that the prediction market generated better predictions than the survey. We also observe that the diversity of beliefs is positively correlated in the survey and the market (Fig. S3). The diversity of beliefs is also higher when the prediction market predicts a low probability that the original result will be replicated. In other words, there is more disagreement about the outcomes of replications that are not likely to be replicated, which could indicate that market participants hold more private information about false positives.

The point-biserial correlation coefficient between the market price and the outcome of the replication is 0.42 and significant ($P = 0.006$, $n = 41$), whereas the survey and weighted survey measures are not significantly correlated with the outcome of the replication [the point-biserial correlation coefficient between the survey and the outcome of the replication is 0.27 ($P = 0.096$, $n = 40$), and the point-biserial correlation coefficient between the weighted survey and the outcome of the replication is 0.26 ($P = 0.112$, $n = 40$)].

Reconstruction of the Prior and Posterior Probabilities p_0 , p_1 , and p_2 from the Market Price p_M

Prior and posterior probabilities associated with the hypothesis are denoted by p_0 , p_1 , and p_2 . Probability p_1 is the prior at the time of the replication, p_2 is the posterior after the replication, and p_0 is the prior at the time of the original study. Probabilities α_0 , β_0 , α_1 , and β_1 are false-positive probabilities and power of the original study and the replication, respectively. Probability p_E denotes the probability of observing positive evidence in the replication, and p_M is the final market price.

From Market Price to p_1 (Eq. 1 in Fig. 2). When the original study reports a positive outcome, successful replication means a positive outcome in the replication. Such a positive outcome can be either due to a true or false positive. The probability p_E for a positive outcome is thus given by $p_E = p_1\beta_1 + (1 - p_1)\alpha_1$. Assuming that the market price p_M reflects probability p_E , probability p_1 can thus be reconstructed as follows:

$$p_1 = (p_M - \alpha_1) / (\beta_1 - \alpha_1). \quad [\text{S1a}]$$

When the original finding is negative, successful replication means a negative outcome in the replication. Thus, the market price p_M reflects $1 - p_E$, rather than p_E , and p_1 is given by the following:

$$p_1 = (1 - p_M - \alpha_1) / (\beta_1 - \alpha_1). \quad [\text{S1b}]$$

From p_1 to p_2 (Eq. 2 in Fig. 2). Once the outcome of the replication is known, it can be used to calculate p_2 from p_1 . In case of a positive outcome, p_2 is given by $p_2 = p_1\beta_1/p_E$. When the original finding is positive, Eq. S1a can be used to substitute p_1 and p_M can be assumed to reflect p_E , and thus p_2 can be calculated as follows:

$$p_2 = (p_M - \alpha_1)\beta_1 / p_M(\beta_1 - \alpha_1). \quad [\text{S2a}]$$

When the original finding is negative, Eq. S1b can be used to substitute p_1 and p_M can be assumed to reflect $1 - p_E$, and thus p_2 can be calculated as follows:

$$p_2 = (1 - p_M - \alpha_1)\beta_1 / (1 - p_M)(\beta_1 - \alpha_1). \quad [\text{S2b}]$$

In case of a negative outcome in the replication, p_2 is given by $p_2 = p_1(1 - \beta_1)/(1 - p_E)$. Thus, in case of a positive original result, p_2 can be calculated as follows:

$$p_2 = (p_M - \alpha_1)(1 - \beta_1) / (1 - p_M)(\beta_1 - \alpha_1). \quad [\text{S2c}]$$

In case of a negative original result and a negative outcome in the replication, p_2 is as follows:

$$p_2 = (1 - p_M - \alpha_1)(1 - \beta_1) / p_M(\beta_1 - \alpha_1). \quad [\text{S2d}]$$

From p_1 to p_0 (Eq. 3 in Fig. 2). Probability p_1 can also be used to reconstruct the original prior, p_0 . When the original result is positive, the original prior is given by the following:

$$p_0 = p_1\alpha_0 / (p_1\alpha_0 + (1 - p_1)\beta_0). \quad [\text{S3a}]$$

When the original result is negative, the original prior is given by the following:

$$p_0 = p_1(1 - \alpha_0) / (p_1(1 - \alpha_0) + (1 - p_1)(1 - \beta_0)). \quad [\text{S3b}]$$

The Association Between the Market Price and the Statistical Power

Based on the section above, one would expect the market price to be positively associated with the statistical power of the original study and the statistical power of the replication. We tested these associations in the data (excluding the study that replicated a null result in the original study). The Pearson correlation coefficient between the market price and the power of the original study is 0.26 ($P = 0.086$, $n = 43$). The Pearson correlation coefficient between the market price and the power of the replication is 0.35 ($P = 0.020$, $n = 43$). In an ordinary least squares regression (with robust SEs) of the prediction market price as a function of the statistical power of the original study and the power of the replication, the R -squared is 14.7% and the regression is significant ($F = 6.56$, $P = 0.003$; both coefficients have the expected signs, but only the coefficient for the power of the replication is significant; $P = 0.027$ for the power of the replication, and $P = 0.321$ for the power of the original study; $n = 43$).

A limitation of these analyses is that there is relatively little variation in the replication power that was constrained to be at least 80% in all replications. In addition, the power of the original studies was estimated ex post based on the P values; ideally ex ante power estimations from the original studies should have been used, but such data were not available from the original studies.

Results and Data for the Individual Studies

In Table S1 (the first set of prediction markets) and Table S2 (the second set of prediction markets), we present the results of p_0 , p_1 , and p_2 for each of the 44 studies included in the prediction market (p_2 could only be estimated for the 41 studies in which the replication has been carried out), along with the data (the market price, the statistical power of the original study, and the statistical power of the replication) used in these estimations. In Table S3 (the first set of prediction markets) and Table S4 (the second set of prediction markets), we report the hypothesis replicated in each study; and in Table S5 (the first set of prediction markets) and Table S6 (the second set of prediction markets), we provide additional data about the prediction markets. The significance level (the false-positive probabilities α_0 and α_1) are set to 5% in all estimations as a significance level

of 5% was used in both the original studies and the replications. The results of the prediction market and the survey are also shown in the tables. For the case of the replication of the originally negative result, we show $1 - p_0$, $1 - p_1$, and $1 - p_2$ in Table S1 and in Fig. 3, because the working hypothesis in the original study in this case was a negative outcome.

For the statistical power of each finished replication study, we use the power of the replication stated in the replicating authors' replication reports. This information is contained on the RPP project page at the Open Science Framework, <https://osf.io/ezcuj/>. For the replications that have not yet been carried out, we use the planned power of the replication also taken from the RPP Open Science Framework project page (which was available information to the prediction market participants at the same location). The statistical power of the original studies was not reported in the published papers. Therefore, we did a post hoc estimate of the statistical power of the original studies based on the P values of the published studies and the standard power formula (i.e., the power estimate is essentially a rescaled P value). This power estimate can be interpreted as the power of finding the observed effect size in the original study at the 5% level with the same sample size as in the original study.

The prediction markets predicted 87% (20 of 23) of the replications correctly in the first set of prediction markets and 50% (9 of 18) of the replications correctly in the second set of prediction markets. These point estimates differ substantially and the prediction rates are significantly different between the two sets of prediction markets ($P = 0.016$; Fisher's exact test). If the prediction market prices are correct estimates of the probability of replication for each individual replication, the expected prediction rate is 69% in the first set of prediction markets and 68% in the second set of prediction markets.

The self-reported expertise about the topic of the studies was significantly lower in the second set of prediction markets compared with the first set of markets (1.71 vs. 1.91, independent-samples t test, $n = 92$, $t = 2.146$, $P = 0.035$). It is possible that this lower self-reported expertise has contributed to less well-functioning prediction markets in the second set of prediction markets. However, the different prediction rates in the first and the second sets of prediction markets may also be due to random variations, especially as the overall prediction rate for the two sets of markets of 71% is close to the expected prediction rate of 69% based on the distribution of market prices.

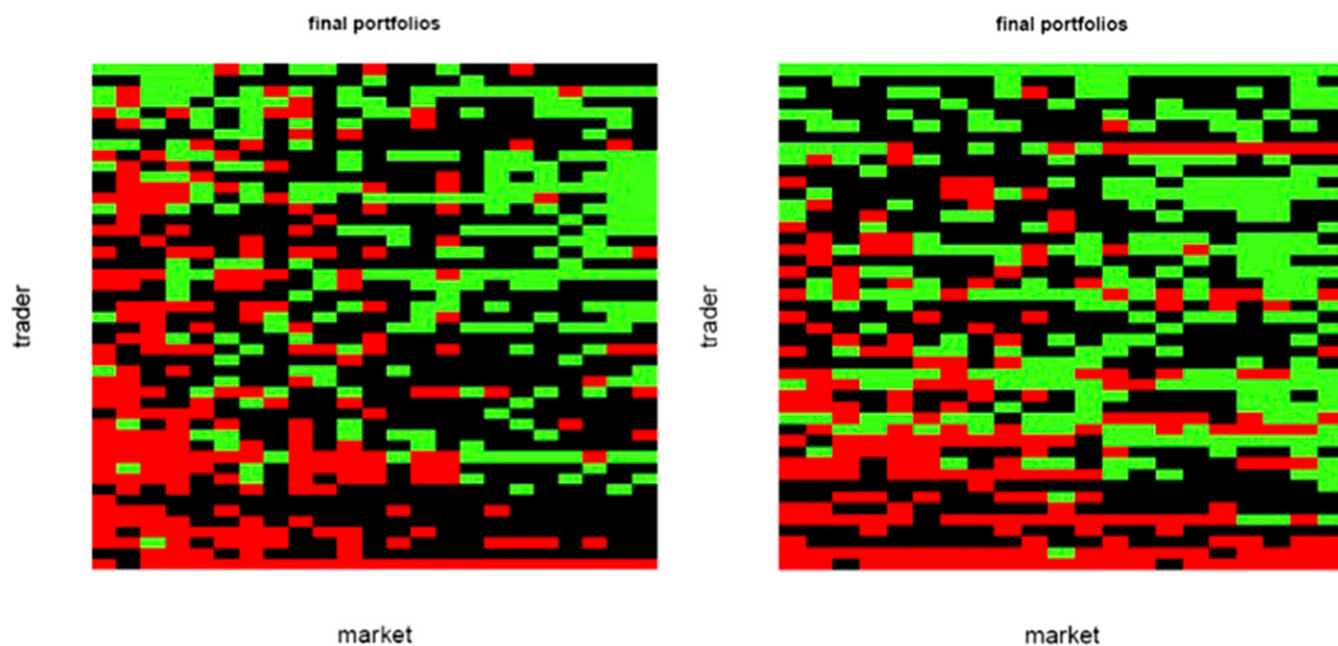


Fig. S1. Final positions per participant and market. The left panel shows the portfolios in the first set of prediction markets, and the right panel shows the portfolios for the second set of prediction markets. Long positions (bets on success) are shown in green, and short positions (bets on failure) are shown in red. This figure indicates that, in both sets of prediction markets, the participants had broad portfolios with positions in several markets. Similarly, each market attracted a number of traders. Often, traders have diverging views: in each market, there is at least one trader holding a long position, and one trader holding a short position. The final portfolios show that there are a few “bears” (predominantly betting on failure) who invested in short positions only (6 of 47 traders for the first set of markets; 4 of 45 traders for the second set of markets), and “bulls” (predominantly betting on success) who invested in long positions only (3 of 47 traders for the first set of markets; 6 of 45 traders for the second set of markets). However, most of the participants fall into a wide spectrum between these two extremes.

A
a

TOPICS

MY ANSWERS

LEADERS

HOW TO PLAY

FAQ

REFER A FRIEND

CONTACT US

DASHBOARD

hypothesis_19
54.76 +4.76

In each of the below questions that you participate, you will bet on a binary outcome: whether or not the replication study finds a statistically significant effect in the same direction as the original study. By statistically significant we mean a p-value smaller than 0.05. By same direction we mean a coefficient that has the same sign as in the original study (i.e. positive or negative).

Hypotheses

	SCORE	
Hypothesis 19, Vul et al., "Temporal selection is suppressed, delayed, and diffused during the attentional blink", Psychological Science (hypothesis_19)	54.76 +4.76	<button>Adjust</button>

 Network 9,998
 My Rank 35
 Available Points 8,998

Hypotheses

TIPS AND TRICKS

A simple range is 500 to 2000 points for each question you wish to answer based on your level of confidence. Be sure to not run out of points.

< ANOTHER TIP >

b

Hypothesis 19, Vul et al., "Temporal selection is suppressed, delayed, and diffused during the attentional blink", Psychological Science

Hypothesis: Attentional selection is suppressed, delayed, and diffused in time during the attentional blink and these effects are dissociated by their time course. Study in paper being replicated: Study no 1 [openscienceframework.org](https://www.openscienceframework.org)

Choose your answer

Yes

No

 **Comments**

Enter your comments here

Post My Comment

c

Enter Your Points

< Your answer: **Yes** Score: 50.00 ?

Points Submit Answer!

Your Points: **9,998** Points After Answer: **8,998**

Fig. S2. (Continued)

B My Answers

Below are the amount of points you currently have invested in each question. Feel free to add or subtract points in a question by clicking the "Adjust" button.

TITLE (SYMBOL)	POSITION	SHARES	VALUE (PROFIT)	
Hypothesis 19, Vul et al., "Temporal selection is suppressed, delayed, and diffused during the attentional blink", Psychological Science (hypothesis_19)	Yes	19	1,083 (+83)	Adjust Close Position
Hypothesis 21, Fischer et al., "Selective exposure and information quantity: How different information quantities moderate decision makers' preference for consistent and inconsistent information", JPSP (hypothesis_21)	Yes	10	368 (-132)	Adjust Close Position
Hypothesis 5, Shnabel, "A needs-based model of reconciliation: Satisfying the differential emotional needs of victim and perpetrator as a key to promoting reconciliation", JPSP (hypothesis_5)	No	38	2,393 (+293)	Adjust Close Position
		Holdings:	3,844	
		Points:	6,398	
		Networth:	10,242	

Fig. S2. (A) Trading interface introductory page. When entering the prediction market, participants were presented with all hypotheses along with their current price ("score") and recent change in price. By clicking Adjust, the participants received more information on the study and the possibility to trade by buying and selling (a). For each replication, participants were presented with the hypothesis, the authors, the title, and the journal, and could buy stocks by choosing Yes or sell stocks by choosing No (b), and enter how many points they would like to invest in the specific hypothesis (c). (B) Position summary presented participants with an overview of their investments: which hypotheses, number of shares held, and current market value.

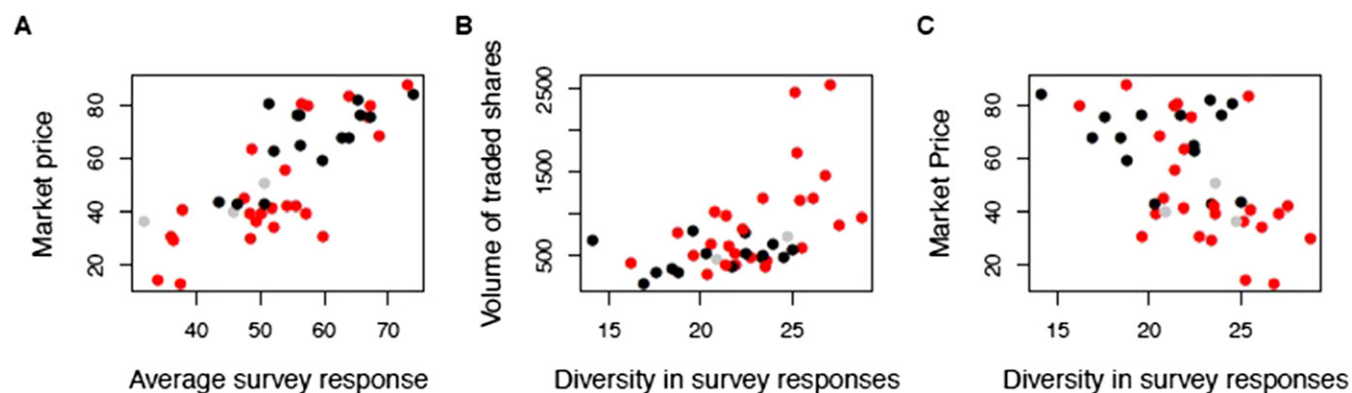


Fig. S3. Comparison of survey responses and behavior in the two prediction markets. (A) Correlation between market price and average survey response. Market prices and average survey responses are positively correlated, suggesting that information given in the surveys was also revealed in the market (Pearson correlation coefficient of 0.78, $P < 0.001$, $n = 43$). However, market prices are more “extreme” than survey responses, which translate into a lower prediction error. Studies that were replicated successfully are shown in black, and studies that failed to replicate are shown in red. Studies that remained unfinished are shown in gray. (B) Correlation between volume of traded shares and diversity in survey responses (i.e., SD of responses; Pearson correlation coefficient of 0.51, $P < 0.001$, $n = 43$). The positive correlation between volume in the market and diversity in the surveys suggests that there was more trading for studies where participants had more diverging views on the replicability of a study. In other words, when there is larger diversity in premarket views, more trades are required to reach a “consensus” in the market pricing. (C) Negative correlation between market price and diversity in survey responses (Pearson correlation coefficient of -0.53 , $P < 0.001$, $n = 43$). The diversity of survey responses is higher when the prediction market predicts a low probability that the original result will be replicated. This suggests that there is more disagreement around replications that are overall expected to fail rather than replications expected to succeed.

Table S1. Individual results for the 23 replication studies in the first set of prediction markets

Ref.	Study no.	Replicated	Market price	Survey result	Weighted survey	Original power	Replication power	p_0	p_1	p_2
33	1	No	36.33	49.39	53.84	0.49	0.91	0.055	0.364	0.051
34	5*	Yes	71.73	.	.	0.9	0.95	0.232	0.741	0.982
35	1 (leadership)	No	29.95	48.49	52.63	0.50	0.56	0.087	0.489	0.307
36	6	No	79.95	67.21	67.74	0.91	0.95	0.215	0.833	0.208
37	4	No	41.77	51.85	53.51	0.73	0.89	0.051	0.438	0.083
38	2	No	30.5	59.8	62.71	0.63	0.86	0.035	0.315	0.063
39	2	No	34.67	52.15	51.49	0.64	0.9	0.040	0.349	0.053
40	2B	No	29.08	36.39	35.06	0.50	0.99	0.033	0.256	0.004
41	2	Yes	83.9	73.96	76.76	0.94	0.99	0.217	0.839	0.990
42	1	No	42.09	55.57	56.63	0.77	0.99	0.041	0.395	0.007
43	2	Yes	64.6	56.22	59.6	0.61	0.99	0.124	0.634	0.972
44	3	Yes	67.37	62.74	63.15	0.99	0.99	0.091	0.664	0.975
45	4b	Yes	67.49	63.85	64.22	0.96	0.99	0.094	0.665	0.975
46	1	No	39.1	57.17	57.81	0.87	0.87	0.039	0.416	0.089
47	1	No	13.22	37.51	35.59	0.60	0.79	0.010	0.111	0.027
48	1	No	75.66	66.8	72.14	0.49	0.94	0.282	0.794	0.196
49	5	No	44.88	47.43	50.73	0.52	0.84	0.089	0.505	0.147
50	4b	Yes	75.35	67.23	68	0.62	0.95	0.224	0.782	0.986
51	1	Yes	76.35	65.66	68.41	0.99	0.99	0.137	0.759	0.984
52	1	No	39.54	50.13	49.89	0.80	0.91	0.040	0.402	0.060
53	4	Yes	59.52	59.72	59.91	0.91	0.99	0.071	0.580	0.965
54	2	No	14.4	33.93	35.17	0.76	0.99	0.007	0.100	0.001
55	3	No	55.82	53.93	52.28	0.59	0.92	0.106	0.584	0.106

*For this study, the authors of the original study hypothesized and found a null effect, and the prediction market was for the prediction that this null result would be replicated.

Ref.	Study no.	Replicated	Market price	Survey result	Weighted survey	Original power	Replication power	p_0	p_1	p_2
56	3	No	83.09	63.87	66.54	0.92	0.9	0.380	0.919	0.543
57	1	No	80.69	56.42	56.05	0.84	0.9	0.326	0.890	0.461
58	2	NA*	39.78	45.8	48.46	0.65	0.82	0.060	0.452	NA
59	5	NA [†]	36.63	31.73	31.31	0.71	0.95	0.037	0.351	NA
60	1	No	41.1	37.78	39.16	0.64	0.93	0.052	0.410	0.049
61	8	No	87.86	73.07	72.93	0.99	0.9	0.662	0.975	0.803
62	5	No	63.35	48.62	49.75	0.62	0.86	0.172	0.720	0.275
63	2	NA*	50.65	50.58	52.30	0.73	0.9	0.074	0.537	NA
64	5	Yes	43.29	43.48	43.59	0.44	0.88	0.089	0.461	0.938
65	1	Yes	76.55	55.8	58.02	0.99	0.99	0.139	0.761	0.984
66	4	Yes	80.32	51.34	54.11	0.58	0.92	0.357	0.866	0.992
67	4	No	39.62	48.31	50.42	0.89	0.99	0.032	0.368	0.006
68	2	Yes	81.59	65.20	68.73	0.48	0.99	0.314	0.815	0.989
69	3	No	30.55	35.98	35.22	0.96	0.95	0.020	0.284	0.020
70	2	Yes	62.6	52.16	51.19	0.94	0.97	0.082	0.626	0.970
71	1	Yes	76.14	56.11	56.76	0.87	0.98	0.158	0.765	0.985
72	1	Yes	42.65	46.34	46.22	0.99	0.99	0.033	0.401	0.930
73	4	No	79.68	57.44	58.54	0.58	0.95	0.296	0.830	0.204
74	1	No	68.25	68.61	71.77	0.99	0.99	0.094	0.673	0.021
75	3	No	42.28	54.16	55.65	0.92	0.99	0.034	0.397	0.007
76	1	Yes	42.89	50.62	52.33	0.99	0.76	0.055	0.534	0.946

¹The replicated experiment from the original paper was changed after the market had already been performed [change from experiment 5 (market) to experiment 1].

Table S3. Hypotheses for the 23 replication studies in the first set of prediction markets

Ref.	Hypothesis
33	White participants with high external motivation to respond without prejudice toward Blacks have an attentional bias toward neutral Black faces presented for 30 ms, but have an attentional bias away from neutral Black faces presented for 450 ms. These biases are eliminated when the faces display happy expressions.
34	Participants do not exhibit a delay in response when switching between pronouncing regular words and pronouncing nonwords.
35	Naive participants' judgments of the power and leadership of CEO faces are correlated positively with their companies' profits.
36	Repetition blindness (a reduction in reporting seeing an orthographically identical or similar word when it is presented in close temporal proximity amid a series of rapidly presented words or nonwords) will occur even for nonidentical orthographical neighbors (e.g., boss and bass) even when the stimuli are nonwords and when they are never repeated in the string of stimuli.
37	An increase in participants' public moral image will be related to an increased willingness to reconcile only for perpetrators, whereas an increase in participants' sense of power will be related to an increased willingness to reconcile only for victims.
38	Participants instructed to avoid race or use race in categorizing tools and guns exhibited less 1/f noise than participants in a control condition where no mention of race was made.
39	Participants with reduced self-regulation resources are expected to exhibit more pronounced confirmatory information processing than nondepleted and ego-threatened participants, whereas no significant differences regarding confirmatory information processing are expected between nondepleted and ego-threatened participants.
40	Participants will prefer descriptions of the city of Los Angeles that are more concrete/less abstract when they are exposed to the words "Los Angeles" during an earlier exercise. Participants who are not shown "Los Angeles" during this earlier exercise will prefer relatively less concrete/more abstract descriptions of the city of Los Angeles.
41	Word processing is slower for dense near semantic neighborhoods, i.e., words with many near neighbors are processed more slowly than words with few near neighbors.
42	Words denoting objects that typically occur high in the visual field hinder identification of targets appearing at the top of the display, whereas words denoting low objects hinder target identification at the bottom of the display.
43	Survival processing yields better memory retention than a control condition with a contextually rich (but non-survival-relevant) encoding scenario.
44	When there are no nonoccurrences of the outcome in the presence of just one cause (cause A), increasing the number of occurrences of the outcome in the presence of that cause alone does not alter the conditional contingency. Under the conditional contingency hypothesis, therefore, such manipulations should not have a significant effect on causal judgment. As opposed to this, the tested predictions are that (i) such occurrences raise judgments of A as cause for the outcome and (ii) lower judgments of an alternative cause B.
45	When participants read sequences of digits and a task requires the joint processing of nonadjacent pairs of digits, they learn exclusively the relation between these nonadjacent digits and not relations between adjacent digits, thus suggesting attention instead of spatial contiguity as the critical factor.
46	Drug use is positively correlated with learning from experience under "sunny" conditions (in which win-loss probabilities are known before making a series of choices) but not correlated under "cloudy" conditions (in which the win-loss probabilities are not known in advance and can only be learned through trial and error).
47	Drinking lemonade with sugar reduces the attraction effect (the reliance on intuitive, heuristic-based decision making) compared with drinking lemonade with sugar substitute among subjects with depleted mental resources.
48	There are semantic interference effects in the delayed naming conditions such that individuals are slower to respond to semantically related word-picture pairs than semantically unrelated word-picture pairs.
49	Participants' ambivalence scores differ across three conditions (implemental mindset one-sided focus, implemental mindset two-sided focus, and neutral mindset), with the implemental mindset one-sided group showing a significantly lower amount of ambivalence compared with the implemental mindset two-sided group. Participants assigned to the neutral mindset condition score in the middle, although not significantly different from either group.
50	Visual statistical learning for colors operates in a feature-based manner if the covariance between feature dimensions is disrupted.
51	Attentional selection is suppressed, delayed, and diffused in time during the attentional blink, and these effects are dissociated by their time course.
52	People who read an essay undermining free will show more cheating in a simple arithmetic task than people who read a control essay.
53	When confronted with more than two pieces of information, the salient selection criterion is expected information quality, which causes a preference for consistent information.
54	There will be a triple interaction with man's availability, participant's conception risk, and participant's partnership status such that man's availability and participant's conception risk interact significantly for partnered women but not for unpartnered ones. In particular, this interaction will show that women with a partner will prefer attached men during the less fertile days of their cycle and single men during the more fertile days of their cycle.
55	When asked to intentionally forget a presented item list, participants will forget items that are repeated twice with several other words in between (spaced presentation) more frequently than when they are not directed to forget. This effect will not occur for items that are repeated twice consecutively (massed presentation).

Ref.	Hypothesis
56	Preschool children and adults who are presented with a 3*3 matrix of color photographs of threat-relevant and threat-irrelevant stimuli and are asked to find the threat relevant target among eight treat-irrelevant distractors or the threat-irrelevant target among eight threat-relevant distractors will detect the threat-relevant target faster than the fear-irrelevant target.
57	Older children select the correct object more frequently than what would be expected by chance in "where" trials.
58	The discontinuity effect (that groups make more competitive choices than individuals) is larger under low partner control-joint control (the low PC-JC matrix) than under high partner control-joint control (the high PC-JC matrix).
59	The effect of processing style on social judgments will be partially mediated by hemisphere activation (stronger right hemisphere activation will be related to assimilation judgments, whereas stronger left hemisphere activation will be related to contrast judgments).
60	Men who feel threatened in their faith of the political and economic climate of their country will show a greater romantic interest in women who are portrayed as embodying benevolent sexist ideals than in women who are portrayed as career oriented, party seeking, active in social cases, or athletic.
61	Participants with higher cognitive ability are better than participants with lower cognitive ability at determining which cards must be turned over to prove the validity of a proposed rule regarding the two sides of four displayed cards.
62	Participants in the "vulnerable condition" will believe that the confederate's expressions were happier than their private feelings, and this effect will be larger in the "vulnerable condition" than in the "control condition."
63	A relationship between self-esteem and later health outcomes is mediated through interpersonal stress.
64	Participants who receive neutral statements will outperform participants who receive feedback expressing self-handicapping thoughts.
65	The contribution rate to pool B in the IPD-MD (intergroup prisoners dilemma-maximizing difference) game will be lower than the contribution rate to pool B in the IPD (intergroup prisoners dilemma) game.
66	When pseudohomophones rather than legal nonwords are used as the nonwords in lexical tasks, there is a stimulus quality and word frequency interaction effect.
67	There is an interaction between commitment (regular vs. first-time donors) and focus intervention (the "to date" condition vs. the "to go" condition). The direction of the interaction is that regular donors donate more in the "to go" than in the "to date" condition and this effect is decreased or reversed for first time donors.
68	Participants faced with the task of selecting the correct ink color for a word presented to them will make more color identification errors in low-contingency trials where words have a low correlation to colors, than in medium-contingency trials where words have a stronger correlation to colors.
69	There is a three-way interaction between response (stereotype consistent vs. stereotype inconsistent), correctness (correct vs. incorrect), and proportion (neutral vs. conflicting).
70	Considering that people sample from a database including a constant ratio of more positive than negative information for three providers, in the two-way interaction effect of provider (manipulating sample size) and valence (positive vs. negative information), the tendency to underestimate the frequency of positive and to overestimate the frequency of negative observations will increase from the provider with the smallest to the provider with the highest overall frequency due to differential regression effects.
71	When participants perform a response time task identifying words as either old or new, response times to misses (old items judged as new) are faster than response times to correct rejections (new items judged as new), indicating that priming can occur independently of recognition.
72	Participants who are expecting to play confrontational video games prefer anger-inducing experiences to exciting experiences.
73	Under low-switch conditions, recall performance of consonants is worse for degraded stimuli compared with normal stimuli.
74	Males show less sensitivity in distinguishing between friendliness and sexual interest than women.
75	The difference in measured "trust and comfort" between the high fairness condition and the low fairness condition is larger for African American participants than for white participants.
76	The sum effect will be more pronounced in the categorically related word pairs priming condition than the misaligned unrelated word pairs priming condition.

Table S5. Additional market data for the 23 replication studies in the first set of prediction markets

Ref.	Replicated	Market price	Volume, no. of traded shares	No. of traders	No. of trades
33	No	36.33	2,461.58	35	101
34	Yes	71.73	785.51	27	61
35	No	29.95	969.94	40	79
36	No	79.95	421.53	27	51
37	No	41.77	530.79	25	44
38	No	30.5	514.03	28	53
39	No	34.67	1,186.66	30	80
40	No	29.08	1,190.72	36	93
41	Yes	83.9	684.82	23	59
42	No	42.09	877.16	29	75
43	Yes	64.6	779.17	28	68
44	Yes	67.37	168.96	18	28
45	Yes	67.49	343.38	22	33
46	No	39.1	278.4	20	34
47	No	13.22	1,464.38	37	100
48	No	75.66	820.94	22	51
49	No	44.88	1,019.52	25	59
50	Yes	75.35	297.08	20	40
51	Yes	76.35	797.31	23	58
52	No	39.54	2,563.77	37	108
53	Yes	59.52	310.52	23	41
54	No	14.4	1,730.51	39	100
55	No	55.82	981.58	22	64

Table S6. Additional market data for the 21 replication studies in the second set of prediction markets

Ref.	Replicated	Market price	Volume, no. of traded shares	No. of traders	No. of trades
56	No	83.09	1,155.33	31	69
57	No	80.69	616.98	27	41
58	NA*	39.78	454.71	22	42
59	NA [†]	36.63	723.49	27	61
60	No	41.1	602.49	29	47
61	No	87.86	788.12	28	57
62	No	63.35	397.29	24	41
63	NA*	50.65	430.52	26	44
64	Yes	43.29	584.83	27	47
65	Yes	76.55	645.74	25	47
66	Yes	80.32	491.62	23	50
67	No	39.62	436.51	27	51
68	Yes	81.59	487.47	28	52
69	No	30.55	489.76	29	47
70	Yes	62.6	530.54	23	44
71	Yes	76.14	376.17	21	43
72	Yes	42.65	505.79	22	49
73	No	79.68	388.62	22	35
74	No	68.25	647.32	31	58
75	No	42.28	364.89	25	43
76	Yes	42.89	527.10	23	48

*This paper had not yet been replicated when the paper was written.

[†]The replicated experiment from the original paper was changed after the market had already been performed [change from experiment 5 (market) to experiment 1].