Open access • Journal Article • DOI:10.1126/SCIENCE.AAF0918

Evaluating replicability of laboratory experiments in economics — Source link

Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck Ho ...+15 more authors

Institutions: California Institute of Technology, Stockholm School of Economics, University of California, Berkeley, National University of Singapore ...+3 more institutions

Published on: 25 Mar 2016 - Science (American Association for the Advancement of Science)

Evaluating replicability of laboratory experiments in Economics

Camerer, Colin and Dreber, Anna and Forsell, Eskil and Ho, Teck-Hua and Huber, Jurgen and Johannesson, Magnus and Kirchler, Michael and Almenberg, Johan and Altmejd, Adam and Chan, Taizan and Heikensten, Emma and Holzmeister, Felix and Imai, Taisuke and Isaksson, Siri and Nave, Gideon and Pfeiffer, Thomas and Razen, Michael and Wu, Hang

California Institute of Technology, Stockholm School of Economics, University of California Berkeley, Berkeley, National University of Singapore, University of Innsbruck, Sveriges Riksbank, New Zealand Institute for Advanced Study, University of Göteborg

3 March 2016

Title: Evaluating replicability of laboratory experiments in economics

Authors: Colin F. Camerer[1,*,†], Anna Dreber[2,†], Eskil Forsell[2,†], Teck-Hua Ho[3,4,†], Jürgen Huber[5,†], Magnus Johannesson[2,†], Michael Kirchler[5,9,†], Johan Almenberg[6], Adam Altmejd[2], Taizan Chan[7], Emma Heikensten[2], Felix Holzmeister[5], Taisuke Imai[1], Siri Isaksson[2], Gideon Nave[1], Thomas Pfeiffer[8], Michael Razen[5], Hang Wu[4]

Affiliations:

[1]California Institute of Technology, 1200 E California Blvd, MC 228-77, Pasadena, CA 91125, USA

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

[3]Haas School of Business, University of California Berkeley, Berkeley, CA 94720-1900, USA

[4]NUS Business School, National University of Singapore, Singapore 119245

[5]Department of Banking and Finance, University of Innsbruck, Universitätsstrasse 15, 6020 Innsbruck, Austria

[6]Sveriges Riksbank, SE-103 37 Stockholm, Sweden

[7]Office of the Deputy President (Research and Technology), National University of Singapore, Singapore 119077

[8]New Zealand Institute for Advanced Study, Private Bag 102904, North Shore Mail Centre, Auckland 0745, New Zealand, and Wissenschaftskolleg zu Berlin - Institute for Advanced Study, D-14193 Berlin, Germany

[9]Centre for Finance, Department of Economics, University of Göteborg, SE-40530 Göteborg, Sweden

*Correspondence to: camerer@hss.caltech.edu

†These first seven authors contributed equally to this work.

Abstract: The reproducibility of scientific findings has been called into question. To contribute data about reproducibility in economics, we replicate 18 studies published in the *American Economic Review* and the *Quarterly Journal of Economics* in 2011-2014. All replications follow predefined analysis plans publicly posted prior to the replications, and have a statistical power of at least 90% to detect the original effect size at the 5% significance level. We find a significant effect in the same direction as the original study for 11 replications (61%); on average the replicated effect size is 66% of the original. The reproducibility rate varies between 67% and 78% for four additional reproducibility indicators, including a prediction market measure of peer beliefs.

One Sentence Summary: In a systematic replication project of experimental studies published in high-impact economics journals 61% replicated.

Main Text:

The deepest trust in scientific knowledge comes from the ability to replicate empirical findings directly and independently, whether through reanalyzing original data or by creating new data. While direct replication of this type is widely applauded (*1*), it is rarely carried out in empirical social science. Replication is now more important than ever, as the reproducibility of results has been questioned in many sciences, such as medicine (*2*-*5*), neuroscience (*6*) and genetics (*7*,*8*). In economics, concerns about inflated findings in empirical (*9*) and experimental analysis (*10*,*11*) have also been raised. In the social sciences, psychology has been the most active in both self-diagnosing the forces creating "false positives", and conducting direct replications (*12*-*15*). Several high-profile replication failures (*16*,*17*) quickly led to changes in journal publication practices (*18*). The recent Reproducibility Project Psychology (RPP) replicated 100 original studies published in three top journals in psychology. The vast majority (97) of the original studies reported "positive findings", but in the replications the RPP only found a significant effect in the same direction for 36% of these studies (*19*).

In this article, we provide insights about how well laboratory experiments in economics replicate. Our sample consists of all 18 between-subject laboratory experimental papers published in the *American Economic Review* and the *Quarterly Journal of Economics* in 2011-2014. The most important statistically significant finding, as emphasized by the authors of each paper, was chosen for replication (see the Supplementary Materials, Section 1 and Tables S1 and S2, for details). We use replication sample sizes with at least 90% power [M=0.92, median(Mdn)=0.91] to detect the original effect size at the 5% significance level. All of the replication and analysis plans were made publicly known on the project website (see the Supplementary Materials, Section 1, for details) and were also sent to the original authors for verification.

There are different ways of assessing replication, with no universally agreed upon "gold standard" (*19*–*23*). We present results for the same replication indicators used in the RPP (*19*). As our first indicator of replication we use a "significant effect in the same direction as in the original study" (though see Gelman & Stern (*20*) for a discussion of the challenges of comparing significance levels across experiments).

The results of the replications are shown in Fig. 1A and Table S1. We find a significant effect in the same direction as the original study for 11 replications (61.1%). This is notably lower than the replication rate of 92% (mean power) that would be expected if all original effects were true and accurately estimated (one-sample binomial test, $P<0.001$).

A complementary method to assess replicability is to test whether the 95% CI of the replication effect size includes the original effect size (*19*) (see Cumming (*21*) for a discussion of the interpretation of confidence intervals for replications). This is the case in 12 replications (66.7%). If we also include the study in which the entire 95% CI exceeds the original effect size, the number of replicable studies increases to 13 (72.2%). An alternative measure, which acknowledges sampling error in both original and replications, is to count how many replicated effects lie in a 95% "prediction interval" (*24*). This count is higher (83.3%) and increases to 88.9% if we also include the replication whose effect size exceeds the upper bound of the prediction interval (See the Supplementary Materials, Section 2, and Fig. S2 for details).

The mean standardized effect size (correlation coefficient, r) of the replications is 0.279, compared to 0.474 in the original studies (see Fig. S3). This difference is significant (Wilcoxon

signed-ranks test, z=-2.98, P=0.003, n=18). The replicated effect sizes tend to be of the same sign as the original ones, but not as large. The mean *relative* effect size of the replications is 65.9%.

The original and replication studies can also be combined in a meta-analytic estimate of the effect size (*19*). As shown in Fig. 1B, in the meta-analysis, 14 studies (77.8%) have a significant effect in the same direction as the original study. These results should be interpreted cautiously as the estimates assume that the results of the original studies do not have publication or reporting biases.

To measure peer beliefs about the replicability of original results, we conducted prediction markets before the 18 replications were done (*25*). Dreber et al. (*26*) suggested this as an additional reproducibility indicator in a recent study presenting evidence for a subset of the replications in the RPP. In the prediction market for a particular target study, peers likely to be familiar with experimental methods in economics could buy or sell shares whose monetary value depended on whether the target study was replicated (see Tables S1 and S2 and Fig. S4). The prediction markets produce a collective market probability of replication (*27*) that can be interpreted as a reproducibility indicator (*26*). The traders' (n=97) survey beliefs about replicability were also collected before market trading to get an additional measure of peer beliefs.

The average prediction market belief is a replication rate of 75.2% and the average survey belief is 71.1% (See Figs. 2 and S5 and Tables S3 and S4 for more details). Both are higher than the observed replication rate of 61.1%, but neither difference is significant (see Supplementary Materials, Section 5, for details). The prediction market beliefs and the survey beliefs are highly correlated, and both are positively correlated with a successful replication, although the correlation does not reach significance for the prediction market beliefs (See Figs. 2 and S6). Contrary to Dreber et al. (*26*) prediction market beliefs are *not* a more accurate indicator of replicability than survey beliefs.

We also test if the reproducibility is correlated with two observable characteristics of published studies: the p-value and the sample size (the number of participants) of the original study. These two characteristics are likely to be correlated with each other, which is also the case for our 18 studies (Spearman correlation=-0.61, P=0.007, n=18). We expect the reproducibility to be negatively correlated with the original p-value and positively correlated with the sample size as the risk of false positives increases with the original p-value and decreases with the original sample size (statistical power) (*6*,*11*). The correlations are presented in Fig. 3 and Table S5, and the results are in line with our expectations. The correlations are typically around 0.5 in the expected direction and significant. Only one study out of eight with a p-value <0.01 in the original study failed to replicate at the 5% level in the original direction.

We report the first systematic evidence of replications of lab experiments in economics, to contribute much-needed data about reproducibility of empirical findings in all areas of science. The results provide provisional answers to two questions: 1) Do laboratory experiments in economics generally replicate? And 2) Do statistical measures of research quality, including peer beliefs about replicability, help predict which studies will replicate?

The provisional answer to question one is that replication in this sample of experiments is generally successful, though there is room for improvement. Eleven out of 18 (61.1%) studies did replicate with P<0.05 in the original direction, and three more studies are relatively close to

being replicated (all have significant effects in the meta-analysis). Four replications (22.2%) have effect sizes close to zero, and those four strong replication failures are somewhat larger in number than the 1.4 expected by pure chance (given the mean power of 92%). Moreover, original effect sizes tend to be inflated which is a phenomenon that could stem from publication bias (*28*). If there is publication bias our prospective power analyses will have overestimated the replication power.

The answer to question two is that peer surveys and market beliefs *did* contain some information about which experiments were more likely to replicate, but sample sizes and p-values in the original studies are even more strongly correlated with replicability (see Fig. 3).

To learn from successes and failures in different scientific fields, it is useful to compare our results with recent results on robustness in experimental psychology and empirical economics.

Our results can be compared to the recent RPP project in the psychological sciences (*19*), which was also accompanied by prediction market beliefs and survey beliefs (*26*). All measures of replication success are somewhat higher for economics experiments than for the sampled psychology experiments (Fig. 4). Peer beliefs in our study are also significantly higher than in the RPP study (Fig. 4). Recognizing the limits of this two-study comparison, and particularly given our small sample of 18 replications, it appears that there is some difference in replication success in these fields. However, it is premature to draw strong conclusions about disciplinary differences; there are other methodological factors that could potentially explain why the replication rates differed. For example, in the RPP replications, interaction effects were less likely to replicate compared to main or simple effects (*19*).

In economics, several studies have shown that statistical findings from non-experimental data are not always easy to replicate (*29*). Two studies of macroeconomic findings reported in the *Journal of Money, Credit and Banking* in 1986 and 2006 could only replicate 13% and 23% of original results, even when data and code were easily accessible (*30,31*). A large analysis of 50,000 reported p-values published between 2005 and 2011 in three widely cited general economics journals shows "missing" p-values between .05-.20 (*32*). However, the frequency of missing values is smaller in lab and field experiments. Taken together, these analyses and our replication sample suggests that lab experiments are at least as robust, and perhaps more robust, than other kinds of empirical economics.

There are two methodological research practices in laboratory experimental economics that may contribute to relatively good replication success. First, experimental economists have strong norms about always motivating subjects with substantial financial incentives, and not using deception. These norms make subjects more responsive and may reduce variability in how experiments are done across different research teams, thereby improving replicability. Second, pioneering experimental economists were eager for others to adopt their methods. To this end, they persuaded journals to print instructions - and even original data - in scarce journal pages. These editorial practices created norms of transparency and made replication and reanalysis relatively easy.

There is every reason to be optimistic that science in general, and social science in particular, will emerge much better off after the current period of critical self-reflection. Our study suggests that lab experimentation in economics published in top journals generates relatively good replicability of results. There are still challenges: For example, executing a few

of the replications was laborious, even when scientific journals require online posting of data and computer code to make things easier. This is a reminder that as scientists we should design and document our methods to anticipate replication and make it easy to do. Our results also show that there is some information in post-publication peer beliefs (revealed in both markets and surveys), and perhaps even more information in simple statistics from published results, about whether studies are likely to replicate. All these developments suggest that cultivation of good professional norms, weeding out bad norms, disclosure requirements policed by journals, and simple evidence-based editorial policies can improve reproducibility of science, perhaps very quickly.

References and Notes:

1. M. McNutt, Reproducibility, *Science* **343**, 229 (2014).
2. J. P. A. Ioannidis, Why most published research findings are false, *PLoS Med* **2**, e124 (2005).
3. F. Prinz, T. Schlange, K. Asadullah, Believe it or not: How much can we rely on published data on potential drug targets?, *Nat. Rev. Drug Disc.* **10**, 712 (2011).
4. C. G. Begley, L. M. Ellis, Drug development: Raise standards for preclinical cancer research, *Nature* **483**, 531 (2012).
5. L. P. Freedman, I. M. Cockburn, T. S. Simcoe, The economics of reproducibility in preclinical research, PLoS Biol **13**, e1002165 (2015).
6. K. S. Button, et al., Power failure: why small sample size undermines the reliability of neuroscience, *Nature Rev. Neurosci.* **14**, 365 (2013).
7. J. K. Hewitt, Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits, *Behavior genetics* **42**, 1 (2012).
8. M. S. Lawrence, et al., Mutational heterogeneity in cancer and the search for new cancer-associated genes, *Nature* **499**, 214 (2013).
9. E. E. Leamer, Let's take the con out of econometrics, *Am. Econ. Rev.* **73**, 31 (1983).
10. A. E. Roth, Let's keep the con out of experimental econ.: A methodological note, *Empir. Econ.* **19**, 279 (1994).
11. Z. Maniadis, F. Tufano, J. A. List, One swallow doesn't make a summer: new evidence on anchoring effects, *Am. Econ. Rev.* **104**, 277 (2014).
12. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant, *Psychol. Sci.* **22**, 1359 (2011).
13. S. Carpenter, Psychology's bold initiative, *Science* **335**, 1558 (2012).
14. Open Science Collaboration, An open, large-scale, collaborative effort to estimate the reproducibility of psychological science, *Perspect. Psychol. Sci.* **7**, 657 (2012).
15. J. Bohannon, Replication effort provokes praise—and 'bullying' charges, *Science* **344**, 788 (2014).
16. S. Doyen, O. Klein, C.-L. Pichon, A. Cleeremans, Behavioral priming: It's all in the mind, but whose mind?, *PLoS ONE* **7**, e29081 (2012).
17. S. J. Ritchie, R. Wiseman, C. C. French, Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect, *PLoS ONE* **7**, e33423 (2012).
18. B. A. Nosek, et al., Promoting an open research culture: author guidelines for journals could help to promote transparency, openness, and reproducibility, *Science* **348**, 1422 (2015).
19. Open Science Collaboration, Estimating the reproducibility of psychological science, *Science* **349** (2015).

20. A. Gelman, H. Stern, The difference between "significant" and "not significant" is not itself statistically significant, *Am. Stat.* **60**, 328 (2006).

21. G. Cumming, Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better, *Perspect. Psychol. Sci*. **3**, 286 (2008).

22. J. Verhagen, E.-J. Wagenmakers, Bayesian tests to quantify the result of a replication attempt, *J. Exp. Psychol.: Gen.* **143**, 1457 (2014).

23. U. Simonsohn, Small telescopes detectability and the evaluation of replication results, *Psychol. Sci.* **26**, 559 (2015).

24. J. T. Leek, P. Patil, R. D. Peng, A glass half full interpretation of the replicability of psychological science, *arXiv* 1509.08968 (2015).

25. K. J. Arrow, et al., The promise of prediction markets, *Science* **320**, 877 (2008).

26. A. Dreber, et al., Using prediction markets to estimate the reproducibility of scientific research, *Proc. Natl. Acad. Sci.* U.S.A. **112**, 15343 (2015).

27. J. Wolfers, E. Zitzewitz, Interpreting prediction market prices as probabilities, Working Paper No. 12200, National Bureau of Economic Research (2006).

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

29. B. D. McCullough, H. D. Vinod, Verifying the solution from a nonlinear solver: A case study, *Am. Econ. Rev.* **93**, 873 (2003).

30. W. G. Dewald, J. G. Thursby, R. G. Anderson, Replication in empirical economics: The journal of money, credit and banking project, *Am. Econ. Rev.* **76**, 587 (1986).

31. B. D. McCullough, K. A. McGeary, T. D. Harrison, Lessons from the JMCB Archive, *J. Money Credit Bank.* **38**, 1093 (2006).

32. A. Brodeur, M. Lé, M. Sangnier, Y. Zylberberg, Star Wars: The empirics strike back, *AEJ: Applied* **8**, 1 (2016).

33. J. Abeler, A. Falk, L. Goette, D. Huffman, Reference points and effort provision, *Am. Econ. Rev.* **101**, 470 (2011).

34. A. Ambrus, B. Greiner, Imperfect public monitoring with costly punishment: An experimental study, *Am. Econ. Rev.* **102**, 3317 (2012).

35. B. Bartling, E. Fehr, K. M. Schmidt, Screening, competition, and job design: Economic origins of good jobs, *Am. Econ. Rev.* **102**, 834 (2012).

36. G. Charness, M. Dufwenberg, Participation, *Am. Econ. Rev.* **101**, 1211 (2011).

37. R. Chen, Y. Chen, The potential of social identity for equilibrium selection, *Am. Econ. Rev.* **101**, 2562 (2011).

38. G. De Clippel, K. Eliaz, B. G. Knight, On the selection of arbitrators, *Am. Econ. Rev.* **104**, 3434 (2014).

39. J. Duffy, D. Puzzello, Gift exchange versus monetary exchange: Theory and evidence, *Am. Econ. Rev.* **104**, 1735 (2014).

40. U. Dulleck, R. Kerschbamer, M. Sutter, The economics of credence goods: An experiment on the role of liability, verifiability, reputation, and competition, *Am. Econ. Rev.* **101**, 526 (2011).

41. K. M. M. Ericson, A. Fuster, Expectations as endowments: evidence on reference-dependent preferences from exchange and valuation experiments, *Q. J. Econ.* **126**, 1879 (2011).

42. E. Fehr, H. Herz, T. Wilkening, The lure of authority: Motivation and incentive effects of power, *Am. Econ. Rev.* **103**, 1325 (2013).

43. D. Friedman, R. Oprea, A continuous dilemma, *Am. Econ. Rev.* **102**, 337 (2012).

44. D. Fudenberg, D. G. Rand, A. Dreber, Slow to anger and fast to forgive: Cooperation in an uncertain world, *Am. Econ. Rev.* **102**, 720 (2012).

45. S. Huck, A. J. Seltzer, B. Wallace, Deferred compensation in multiperiod labor contracts: An experimental test of Lazear's model, *Am. Econ. Rev.* **101**, 819 (2011).

46. J. Ifcher, H. Zarghamee, Happiness and time preference: The effect of positive affect in a random-assignment experiment, *Am. Econ. Rev.* **101**, 3109 (2011).

47. J. B. Kessler, A. E. Roth, Organ allocation policy and the decision to donate, *Am. Econ. Rev.* **102**, 2018 (2012).

48. M. Kirchler, J. Huber, T. Stöckl, Thar she bursts: Reducing confusion reduces bubbles, *Am. Econ. Rev.* **102**, 865 (2012).

49. S. Kogan, A. M. Kwasnica, R. A. Weber, Coordination in the presence of asset markets, *Am. Econ. Rev.* **101**, 927 (2011).

50. I. Kuziemko, R. W. Buell, T. Reich, M. I. Norton, "Last-place aversion": Ev-idence and redistributive implications, *Q. J. Econ.* **129**, 105 (2014).

51. B. Merlob, C. R. Plott, Y. Zhang, The CMS auction: experimental studies of a median-bid procurement auction with nonbinding bids, *Q. J. Econ.* **127**, 793 (2012).

52. C. C. Eckel, R. Petrie, Face value, *Am. Econ. Rev.* **101**, 1497 (2011).

53. D. Gill, V. Prowse, A structural analysis of disappointment aversion in a real effort competition, *Am. Econ. Rev.* **102**, 469 (2012).

54. N. Erkal, L. Gangadharan, N. Nikiforakis, Relative earnings and giving in a real-effort experiment, *Am. Econ. Rev.* **101**, 3330 (2011).

55. U. Fischbacher, z-Tree: Zurich toolbox for ready-made economic experiments, *Exp. Econ.* **10**, 171 (2007).

56. S. Palan, GIMS—Software for asset market experiments, *J. Behav. Exp. Fin.* **5**, 1 (2015).

57. R. Hanson, Could gambling save science? Encouraging an honest consensus, *Soc. Epistemol.* **9**, 3 (1995).

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

59. J. Wolfers, E. Zitzewitz, Prediction markets, *J. Econ. Perspect.* **18**, 107 (2004).

60. G. Tziralis, I. Tatsiopoulos, Prediction markets: An extended literature review, *J. Pred. Markets* **1**, 75 (2007).

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

62. C. F. Horn, B. S. Ivens, M. Ohneberg, A. Brem, Prediction markets – a literature review 2014, *J. Pred. Markets* **8**, 89 (2014).

63. C. F. Manski, Interpreting the predictions of prediction markets, *Econ. Letters* **91**, 425 (2006).

64. U. Sonnemann, C. F. Camerer, C. R. Fox, T. Langer, How psychological framing affects economic market prices in the lab and field, *Proc. Natl. Acad. Sci.U.S.A.* **110**, 11779 (2013).

65. R. Hanson, Logarithmic market scoring rules for modular combinatorial information aggregation, *J. Pred. Markets* **1**, 3 (2007).

66. Y. Chen, Markets as an information aggregation mechanism for decision support, Doctor of Philosophy Thesis, School of Information Sciences and Technology, The Pennsylvania State University (2005).

Acknowledgments: For financial support we thank: Austrian Science Fund FWF (START-grant Y617-G11), Austrian National Bank (grant OeNB 14953), Behavioral and Neuroeconomics Discovery Fund (CFC), Jan Wallander and Tom Hedelius Foundation (P2015-0001:1 and P2013-0156:1), Knut and Alice Wallenberg Foundation (Wallenberg Academy Fellows grant to A. Dreber), Swedish Foundation for Humanities and Social Sciences (NHS14-1719:1), and Sloan Foundation (G-2015-13929). We thank the following experimental labs for kindly allowing us to use them for replication experiments: Center for Behavioral Economics at National University of Singapore, Center for Neuroeconomics Studies at Claremont Graduate University, Frankfurt Laboratory for Experimental Economic Research, Harvard Decision Science Laboratory, Innsbruck ECONLAB, and Nuffield Centre for Experimental Social Sciences. We thank the following persons for assistance with the experiments: Agneta Berge, Rahul Bhui, Andreas Born, Nina Cohodes, Ho Kinh Dat, Christoph Dohmen, Zayan Faiayd, Malte Heissel, Austin Henderson, Gabe Mansur, Jutta Preussler, Lukas Schultze, Garrett Thoelen, and Elizabeth Warner. The data reported in this paper are tabulated in Tables S1-S3 and the Replication Reports, analyses code, and the data from the replications are available at www.experimentaleconreplications.com and at OSF (osf.io/bzm54). The authors report no potential conflicts of interest. No MTAs, patents or patent applications apply to methods or data in the paper.

Author Contributions: CC, AD, JH, TH, MJ, and MK designed research; CC, AD, EF, JH, TH, MJ, and MK wrote the paper; EF, JA, TC, TH, TP helped design the prediction market part; EF, FH, JH, MK, MR, TP, and HW analysed data; AA, EH, FH, TI, SI, GN, MR, and HW carried out the replications (including re-estimating the original estimate with the replication data); all authors approved the final manuscript.

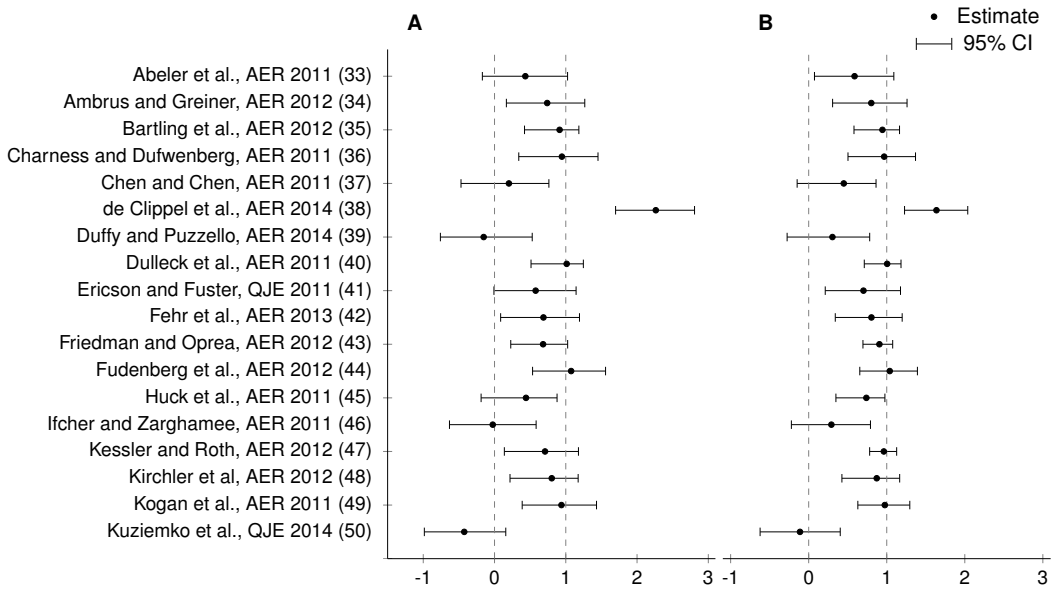

Fig 1. Replication results. (**A**) Plotted are 95% CIs of replication effect sizes (standardized to correlation coefficients r). The standardized effect sizes are normalized so that 1 equals the original effect size (see Fig. S1 for a non-normalized version). There is a significant effect in the same direction as in the original study for 11 replications [61.1%; 95% CI =(36.2%, 86.1%)]. The 95% CI of the replication effect size includes the original effect size for 12 replications [66.7%; 95% CI =(42.5%, 90.8%)]; if we also include the study in which the entire 95% CI exceeds the original effect size, this increases to 13 replications [72.2% [95% CI =(49.3%, 95.1%)]. AER denotes the *American Economic Review* and QJE denotes the *Quarterly Journal of Economics*. (**B**) Meta-analytic estimates of effect sizes combining the original and replication studies. 95% CIs of standardized effect sizes (correlation coefficient r). The standardized effect sizes are normalized so that 1 equals the original effect size (see Fig S1 for a non-normalized version). Fourteen studies have a significant effect in the same direction as the original study in the meta-analysis [77.8%; 95% CI =(56.5%, 99.1%)].

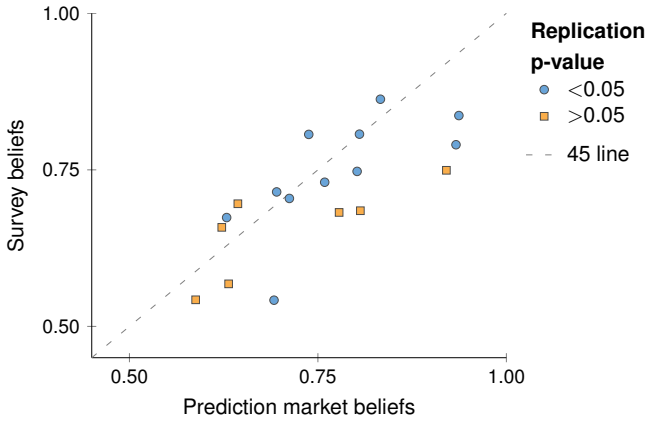

Fig. 2. Prediction market and survey beliefs. A plot of prediction market beliefs and survey beliefs in relation to if the original result was replicated with P<0.05 in the original direction. The mean prediction market belief is 75.2% [range 59% to 94%, 95% CI=(69.7%, 80.6%)], and the mean survey belief is 71.1% [range 54% to 86%, 95% CI =(66.4%, 75.8%)]. The prediction market beliefs and survey beliefs are highly correlated (Spearman correlation coefficient 0.79, P<0.001, n=18). Both the prediction market beliefs (Spearman correlation coefficient 0.30, P=0.232, n=18), and the survey beliefs (Spearman correlation coefficient 0.52, P=0.028, n=18) are positively correlated with a successful replication.

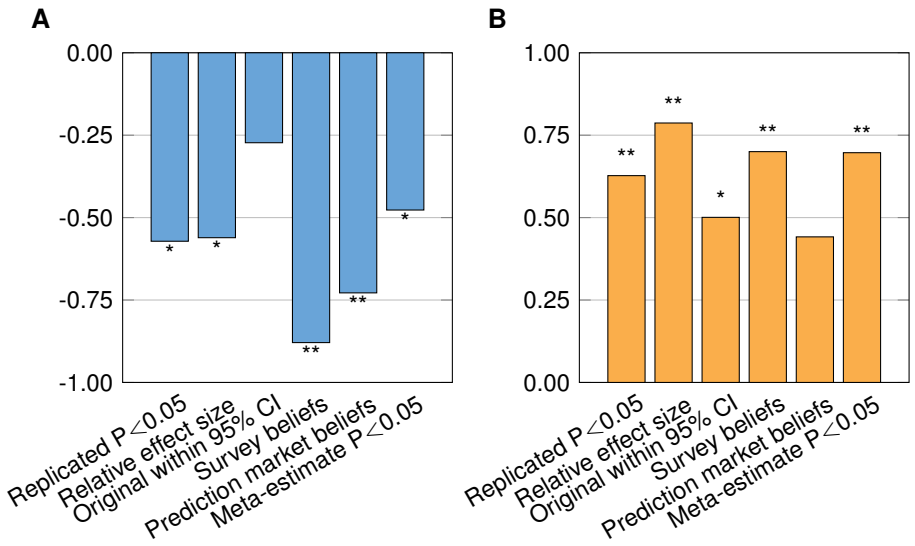

Fig. 3. Correlations between original study p-value and N and reproducibility indicators. The original p-value is negatively correlated with all six reproducibility indicators, and five of these correlations are significant. The original sample size is positively correlated with all six reproducibility indicators, and five of these correlations are significant. Spearman correlations; $^{*}P<0.05$, $^{**}P<0.01$.

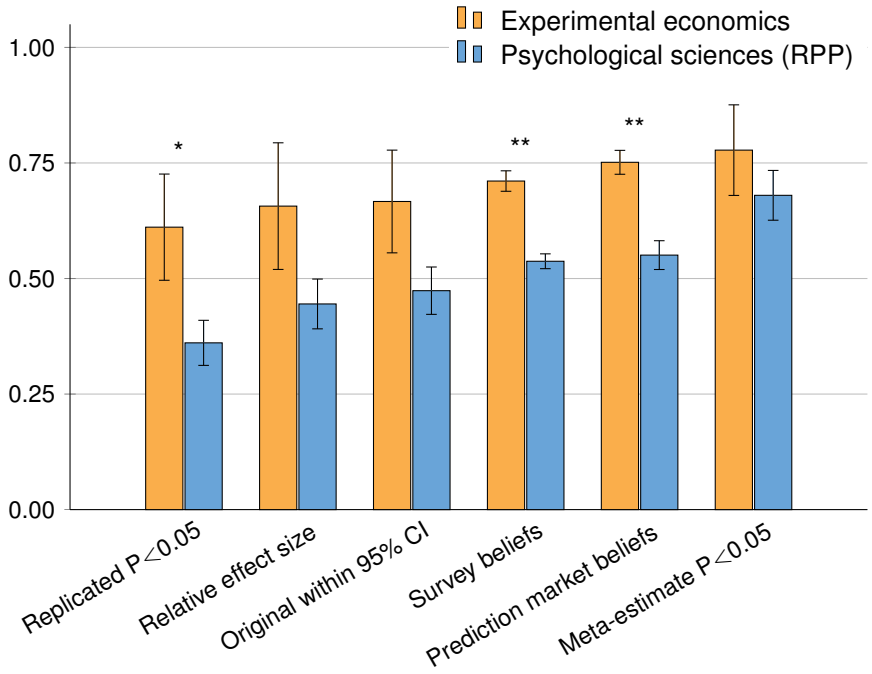

Fig. 4. A comparison of different reproducibility indicators between experimental economics and psychological sciences (the Reproducibility Project Psychology). Error bars denotes ±se. The reproducibility is higher for experimental economics for all six reproducibility indicators; this difference is significant for three of the reproducibility indicators. The average difference in reproducibility across the six indicators is 19 percentage points. See the Supplementary Materials for details about the statistical tests. $^{*}P<0.05$ for the difference between experimental economics and psychological sciences, $^{**}P<0.01$ for the difference between experimental economics and psychological sciences.

Supplementary Materials for

Do lab experiments in economics replicate?

Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, Hang Wu

correspondence to: camerer@hss.caltech.edu

This PDF file includes:

Materials and Methods

Here we provide further details on the replications (Section 1), the estimation of standardized effect sizes and meta-analysis (Section 2), the implementation of the prediction markets and survey (Section 3), the prediction markets performance (Section 4), the comparison of prediction market beliefs and survey beliefs (Section 5), the comparison of reproducibility indicators between experimental economics and psychological sciences (Section 6), and additional results and data for the individual studies/markets (Section 7).

1. Replications

We replicate 18 experimental studies published between 2011 and 2014 in the high-impact general interest journals the *American Economic Review* (AER) and the *Quarterly Journal of Economics* (QJE). (*33–50*) The deadline for inclusion in the study was that the paper should be published or posted as accepted/in press at the website of the journal at August 1, 2014.

There are a number of different possible experimental designs. The most "classical" design is the randomized controlled trial (RCT) design, where participants are randomly allocated to two or more treatments and the outcome is compared between treatments. This design is for instance the gold standard in medicine in comparing different medical treatments. The RCT is a between subjects treatment comparison, and this design is also commonly used in experimental economics (although it is not always the case that participants are strictly randomly allocated to treatments). Another commonly used design is a within subject treatment comparison where the same participants are exposed to two or more treatments and the outcome is compared between the treatments. A within subject design is typically considered a somewhat weaker identification of treatment effects as being exposed to the first treatment may affect behavior in the second treatment. A third common design in experimental economics is to compare the behavior of a group of participants with a theoretical prediction (e.g. to test if behavior in the dictator game is consistent with money maximizing behavior).

In this study we decided to include all between subject treatment comparison studies for replication. To be part of the study a published paper needed to report at least one significant between subject treatment effect that was referred to as statistically significant in the paper, and was emphasized as an important finding by the authors of the paper (e.g. highlighted in the Abstract or the Introduction). If a paper reported more than one significant between subject treatment effects, we used the following 4 criteria in descending order to determine which treatment effect to replicate.

1. The most central result in the paper (among the between subject treatment comparisons) based on to what extent the results were emphasized in the published papers.
2. If more than one equally central result, we picked the result (if any) related to efficiency, as efficiency is central to economics.
3. If several results still remained and they were from different separate experiments we followed the procedure used in the Reproducibility Project Psychology (*19*) and picked the last experiment.

4. In case several results still remained we randomly selected one of those results for the replication. This happened for five studies (*38-40*,*49*,*50*).

If an original study included more than two within subject treatments we only replicated the two treatments used for the result selected for replication. We excluded papers that already included a replication in another subject population; one paper was excluded for this reason (*51*). We also excluded papers that were replications of previous studies; one study was excluded for this reason (*11*). We furthermore excluded studies focusing on interaction effects with treatments; two studies were excluded for this reason (*52,53*) and studies where participants were selected into treatments based on performance in the experiment (one study (*54*) was excluded for this reason).

There were some borderline cases. The study by Fehr et al. (*42*) was included, despite mainly being a within subject treatment study (but it also included a between subjects treatment comparison emphasized by the authors). The Kuziemko et al. (*50*) study was included although the treatment effect was estimated based on both between and within subject treatment variation.

There were four replication teams: a team at Stockholm School of Economics (responsible for 5 replications); a team at University of Innsbruck (responsible for 5 replications), at team at CalTech (responsible for 4 replications), and a team at University of California Berkeley/National University of Singapore (responsible for 4 replications). Replications were not always conducted at the universities of the teams (other labs were also used). Five out of the 18 original experiments were conducted in German, and the remaining ones in English. The 5 original experiments in German were replicated in German speaking populations. Eleven out of the 13 original experiments in English were replicated in English and the remaining two studies were replicated in German. The same software and computer programs as in the original experiments were used to conduct the replications, with the exception of the replication of Kogan et al. (*49*) where the replication was conducted with z-Tree (*55*) and GIMS (*56*) instead of the original software (as the software used by Kogan et al. (*49*) was an online application, which was no longer maintained and therefore impossible to use).

The replication team responsible for each replication wrote a Replication Report detailing the planned replication (with the following sections: Hypothesis to bet on, Power analysis, Sample, Materials, Procedure, Analysis, Differences from original study). A draft of the Replication Report was sent to the original authors for comments, and the Replication Reports were revised based on the comments and then posted at www.experimentaleconreplications.com (we also saved all communications between the original authors and the replication teams on a special e-mail account). After the replications had been conducted the Replication Reports were updated with the results of the replication (the following three sections were added to the reports: Results, Unplanned protocol deviations, Discussion). After all replications had been completed, the Replication Reports were again sent to the original authors for comments. After a revision the final versions were then posted at www.experimentaleconreplications.com (both the versions prior to the replications and the final versions are posted and publicly available).

Everyone involved with carrying out the replications did not receive any information about the prediction markets results or the survey results until all replications had been conducted. Only three members of the research team (Eskil Forsell and Thomas Pfeiffer,

and programmer Taizan Chan) had access to information about the prediction market results prior to the completion of the replications. Those three people were not involved in any replication data collections. Everyone involved with carrying out the replications were also instructed not to discuss the prediction market with any of the individuals who participated in the prediction market. This was done to rule out that the persons conducting the experiments were affected by the prediction market results in carrying out the replications.

All replications were carried out with at least 90% statistical power. In some cases the statistical power was larger than 90% depending on the group sizes used in the experiments. For example, if a group size of 8 subjects was used in the original experiment, and with randomization to two treatments within each session, the total sample size used in the replication needed to be evenly divisible by 16. Subjects were randomly allocated to the two treatments in all replications (even if this was not done in the original experiment; in the original experiments it is sometimes unclear if participants were randomly allocated to treatments or not). If possible we randomly allocated subjects to the two treatments within each session to control for any session/experimenter/time of day effects. In some cases this was not possible due to restrictions on the number of participants in the lab at the same time.

The sample size needed for 90% statistical power to detect the same effect size as in the original study was estimated in the same way for all the replications. We estimated the fraction of the original sample size needed to get 90% power based on the standard power formula of a z-test. This fraction is given by: $(3.242/z)^2$; where z is the z-value in the original study. This formula was used also for studies not using a z-test. In these cases the reported p-value in the study was converted to the corresponding z-value and then the above formula was applied. The power estimation for these studies is thus an approximation.

2. Estimation of standardized effect sizes and meta-analysis

To compare the effect size between the original study and the replication study we transformed effect sizes into correlation coefficients (r) in the same way as done for the RPP project (*19*). Apart from being a well known and bounded effect size measure, the standard errors of the correlation coefficients are very easy to calculate by applying the Fisher transformation and depend only on the sample size of the study (with the sample size here defined as the number of sessions rather than the number of participants if the test is based on session averages, and the number of clusters rather than the number of participants if the test is based on regressions with clustered standard errors). We coded the correlation coefficient to be positive for the original study regardless of the actual sign to allow negative coefficients from the replication studies to be interpreted as going in the opposite direction from the original. The relationship between the original and replication standardized effect sizes (r) can be seen in Fig. S3.

For each study-pair we also computed a fixed-effect weighted meta-analytic effect size measure as also done for the RPP project (*19*). This meta-analytic effect size treats original and replicated studies equally (except for sample size) and represents the best inference of effect size when the studies are taken together. More details about these calculations and the code can be found at www.experimentaleconreplications.com.

We also used the estimated standardized effect sizes to carry out an estimation of replicability with a "small telescopes" approach recently proposed (*23*). The approach entails testing if the replication obtains an effect that is significantly smaller (with a one-sided test at the 5% level) than a "small effect" in the original study, where a small effect is defined as the effect size the original study would have had 33% power to detect. We use an adaptation of the R-package "pwr" (available at www.experimentaleconreplications.com) to calculate these "small effects". If the replication obtains an effect that is significantly smaller than a "small effect" with this definition it is considered a failed replication. For our study this approach yields identical results to the meta-analyses (with the same four studies failing to replicate as in the meta-analyses).

Another approach recently proposed by Leek, Patil & Peng (*24*), is to estimate a 95% prediction interval for the original estimate and test how many of the replications that fall within this prediction interval. We did this estimation as well and 15 replications (83.3%) are within the 95% prediction intervals (Fig. S2); if we also include the replication with an effect size larger than the upper bound of the prediction interval this increases to 16 replications (88.9%). This can be compared to the estimations for RPP by Leek, Patil & Peng (2015); they found that 75% of the replications were within the prediction intervals and 77% if they included the two replications with effect sizes larger than the upper bound of the prediction intervals.

We use the standardized effect sizes (r) to compare results between the original and the replication study and to estimate a measure of the relative effect size of the replication. However, it should be noted that caution has to be exercised in comparing the levels of the standardized effect sizes (r) between the 18 studies. The reason for this is that the aggregation level used in the statistical tests in the original studies varies between studies. Several studies carry out the tests based on session averages (so that one session average becomes one observation in the statistical test), and aggregating the data on the session level reduces the variance of the data (i.e. the variance between individuals is larger than the variance between session averages). A higher degree of aggregation of the data and thus lower variance generally increases the standardized effect size (i.e. the same treatment difference will result in different standardized effect sizes depending on if the statistical test is carried out at the individual level or the session level). A similar issue arises for studies clustering the standard error on, for instance, the session level (where the number of degrees of freedoms will then be based on the number of clusters rather than the number of individuals included in the analysis). For comparing the standardized effect size between the original study and the replication these concerns about sensitivity of how effect sizes are computed is not a problem for inference about replicability, because the statistical tests are carried out in an identical way (the same level of aggregation) in both the original study and the replication.

Due to the limited comparability of the standardized effect sizes between the 18 studies, we present the replication results after normalizing the original effect size to 1 in Fig. 1 (i.e. the upper and lower bound of the 95% CI of the standardized effect (r) in the replication is divided by the standardized effect size (r) in the original study). For this reason we also refrain from analysing the correlation between the original effect size and the reproducibility indicators (although this correlation was reported in the RPP project (*19*). In Fig. S1 we include a non-normalized version of Fig. 1.

In Table S1 we also include an estimate of the relative effect size based on unstandardized effect sizes. This measure is created by dividing the absolute treatment difference in the replication study by the absolute treatment difference in the original study (see the Replication Reports for more details). This relative effect size measure is highly correlated with the relative effect size measure based on standardized effect sizes (the Spearman correlation between the two measures is 0.90 ($P<0.001$)).

3. Implementation of prediction markets and surveys

Prediction markets can aggregate private information on reproducibility, and can generate and disseminate a consensus among market participants. Hanson (*57*) first suggested that prediction markets could be a potentially important tool for assessing scientific hypotheses. Almenberg et al. (*58*) conducted a lab-based test. More recently a prediction market study on replications in psychology has yielded promising results (*26*), in the sense that predictions revealed by market prices are correlated with actual replication outcomes (and are more strongly correlated than surveyed beliefs). Prediction markets have also been successfully used in several other fields such as sports, entertainment, and politics (*25*,*59-62*).

For each of the 18 replication studies we implemented a prediction market where shares whose value was determined by the outcome of the replication could be traded. To be able to relate the performance of the markets to more traditional belief elicitation we also implemented two surveys, one before the markets opened and one after they had closed.

Invitations to participate in the prediction markets were sent to the Economic Science Association mailing list and members of the Editorial Board of the following economics journals: American Economic Review, Quarterly Journal of Economics, Review of Economic Studies, Econometrica, Journal of Political Economy, Experimental Economics, Journal of Economic Behavior and Organization, and Games and Economic Behavior. Authors of the 18 original studies to be replicated were excluded from participating, as was anyone studying at a masters level or lower. The invitation email contained a link to a form where participants could sign up using their university email address.

The number of participants at each stage was as follows: 177 individuals originally signed up to participate; 140 of these filled in the pre-market survey; 97 participated on the prediction markets; and 79 participated in a post-market survey. The number of traders active in each of the markets ranged from 31 to 68. The two largest groups of participants were PhD students and PostDocs (34.4% and 19.8% respectively) and a substantial share held a professor's title of some sort (40.2%). Of the latter group, just under half held a full professor's title (46.2%). Among those participants who stated the time spent in academia (77.3% did so), the average was 7 years (SD=0.853). A majority of participants resided in Europe (54.6%) and the second largest group resided in North America (30.9%).

Invitations to participate were e-mailed on April 2nd 2015, the pre-market survey had to be completed before the 20th for the participant to be invited to the markets, the markets opened on the 22nd, the markets closed on May 3d and the post-market survey had no completion deadline.

The pre- and post-market surveys (available at www.experimentaleconreplications.com) were designed to elicit the same type of information as the prediction markets. In the pre-market survey participants were for each replication study asked to assess: 1) the likelihood that the hypothesis would be replicated; 2) the final trading price in the markets; 3) their stated expertise for the study the hypothesis was taken from; and 4) their confidence of their answer for the first two questions. Participants could also optionally answer a few demographic questions. In the post-market survey participants were again asked to answer questions 1) and 4). The survey questions were not incentivized.

To implement the prediction markets we designed our own web based trading platform. The trading interface contained two main views: 1) the market overview and 2) the trading page. The market overview showed the 18 markets along with some summary information and a trade button for each market (see Fig. S4A). Clicking the trade button for a market showed the trading page where the participant could make investment decisions and view more detailed information about the market (see Fig. S4B).

Participants were endowed with 100 Tokens when the markets opened. These Tokens could be used to trade shares in the markets. If a study replicated (according to the criteria of a significant effect in the same direction as in the original study) shares in the market corresponding to that study were worth one Token each, zero otherwise. This type of contract can under some conditions be interpreted as the average predicted probability of the outcome occurring (*27*,*63*); see Sonneman et al (*64*) for lab evidence that averaged beliefs are close to prediction market prices. All markets opened at a price of 0.50 Tokens per share and were thereafter determined by a market-maker implementing a logarithmic market scoring rule (*65*). The market maker calculates the price of a share for each infinitesimal transaction and updates the price according to the scoring rule. This ensures both that trades are always possible even when there is no other participant with whom to trade and that participants have incentives to invest according to their beliefs (*66*).

The logarithmic scoring rule uses the net sales (shares held - shares borrowed) the market maker has done so far in a market to determine the price for an (infinitesimally small) trade as $p=e^{s/b}/(e^{s/b}+1)$. 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 which meant that by investing 10 Tokens (i.e. 1/10 of the initial endowment), traders could move the price of a single market from 0.50 to about 0.55; and investing the entire initial endowment into a single market moved the price from 0.50 to 0.82.

Investment decisions for a market were made from the market's trading page. Participants could see the (approximate) price of a new share, the number of shares they currently held and the number of Tokens their current position was worth if they liquidated their shares. The trading page also contained information about previous price and aggregate long and short positions presented as graphs. To make an adjustment to their current position participants could choose either to increase or decrease their position by a number of Tokens of their choice. Depending on their current position these actions could have different outcomes.

Increasing a position when holding zero or more shares was equivalent to purchasing new shares at the current price. Decreasing a position when having shorted

zero or more shares was equivalent to short selling new shares at the current price. The repurchasing cost of shorted shares was withheld from the participant's account to ensure that the participant would have enough Tokens to return the shorted shares if the study replicated. For example: a share shorted at a market price of 0.60 Tokens immediately awarded the participant 0.60 Tokens but also stood the risk of having to be bought back at 1 Token if the study replicated. To make sure that the participant could buy back the share in this worst-case scenario, 1 Token was withheld from the participant's account resulting in a deduction of 0.40 Tokens (0.60 Tokens - 1 Token). This setup did not disproportionately discourage short selling as the deducted amount is analogous to the price paid when going long.

Decreasing a position by a moderate amount when already holding shares was equivalent to selling a number of shares. Increasing a position by a moderate amount when having shorted shares was equivalent to buying and returning a number of shares and receiving the withheld Tokens.

If the adjustment to a position was large enough one of the last two outcomes could be combined with one of the first two. Decreasing a position could result in a participant selling all shares they currently held in the market as well as short selling additional shares. Increasing a position could similarly result in a participant returning all shorted shares in the market as well as buying additional shares.

The markets were resolved after all replications were finished. If a replication was successful, shares held in the corresponding market were worth 1 Token each and the Tokens withheld for shorted shares were not returned. If a replication was unsuccessful, shares held in the corresponding market were worth nothing and Tokens withheld for shorted shares (1 Token / share) were returned. Tokens awarded as a result of holding or having shorted shares were converted to USD at a 0.5 rate but Tokens that had not been invested in a market were not converted at all.

To aid their investment decisions all participants had access to the Replication Reports for each replication (the version of the Replication Reports before the replications were conducted), and the references to the original papers. For each replication study participants were informed about the hypothesis to be replicated, the p-value of the original result and the sample size and statistical power. The statistical power was at least 90% to replicate the original effect size at the 5% level.

Investments were settled in the beginning of 2016 according to actual results of the replications.

The prediction market methodology used in this study is similar to the one used in Dreber et al. (*26*). Dreber et al. (*26*) presented prediction market results for 44 studies in RPP. The trading platforms and the participant pool differed between the two prediction market studies (a sample of psychologists participated in the Dreber et al. (*26*) study and a sample of economists, see above, participated in this study).

4. Prediction market performance

The mean trading volume on the prediction markets in terms of traded shares was 1541.1 (median=1458.0) with a range between 733.9 and 2849.4, and in terms of tokens the mean was 507.1 (median=473.0) and the range was 254.7-946.5.

We can distinguish between 6 types of investments; only buying shares, only selling shares, only shorting shares, only returning shares, returning and buying shares, and

selling and shorting shares. The total number of transactions was 1073 for “buy only”, 120 for “sell only”, 427 for “short only”, 387 for “return only”, 36 for “return and buy”, and 37 for “sell and short”.

Fig. S5 shows an overview of market thickness, trader diversification and general trends in shares held and borrowed across all markets and participants.

5. Comparison of prediction market beliefs and survey beliefs

To compare the survey results to the prediction markets results we base the pre-market survey measure on the sample of individuals who participated on the prediction markets (n=97). This is the measure referred to as “survey beliefs” in the main text. But for completeness we also include data for all 140 individuals who completed the pre-market survey in Table S3 (below we also briefly mention how using survey data for all these 140 individuals affect the survey results). The pre-market survey and prediction markets results are quite strongly related (Fig. 2). The Spearman correlation between the prediction market beliefs (final market prices) and the pre-market survey is 0.79 ($P<0.001$, n=18). The range of predictions in the markets is 59%-94% with a mean of 75.2% as compared to the survey range of 54% to 86% and a mean of 71.1%. This is higher than the observed replication rate of 61%, but this difference is not significant for the prediction market beliefs (Wilcoxon signed-ranks test, n=18, z=0.85, P=0.396) or the survey beliefs (Wilcoxon signed-rank test, n=18, z=0.85, P=0.396); note that both these tests produce the same test statistics in spite of the somewhat higher mean prediction market beliefs as they are based on ranks.

One way of evaluating how well the prediction market beliefs and survey beliefs predict the replication outcomes is to interpret a market belief (survey belief) larger than 50% as predicting successful replication and a market belief (survey average) below 50% as predicting failed replication (a successful replication is here and in the analyses below defined as a statistically significant (at the 5% level) effect in the same direction as in the original study; but in Table S5 we also provide the correlations between market and survey beliefs and the other replication indicators). Informative markets are expected to correctly predict more than 50% of the replications. However, as all market beliefs and all survey beliefs are above 50% in this study, the correct prediction rate with this criteria will simply be the replication rate of 61.1% [95% CI =(36.2%, 86.1%)]. This can be compared to the expected replication rate of 75% for the prediction market and 71% for the survey.

The Spearman (rank-order) correlation coefficient between the market beliefs and the outcome of the replication is 0.30, but it is not significant (P=0.232, n=18). The Spearman correlation coefficient between the pre-market survey beliefs and the outcome of the replication is 0.52 (P=0.028, n=18). The absolute prediction error does not differ significantly between the prediction market (Mean=0.414) and the pre-market survey (Mean=0.409) (Wilcoxon signed-rank test, n=18, z=0.33, P=0.744). Contrary to a recent prediction market study on a subset of the studies (n=44) included in the RPP project (*26*), the prediction market thus does not predict replication outcomes better than the survey. However, the sample size of replications is small with only 18 observations. If we average the market beliefs and the pre-market survey beliefs the Spearman correlation with the outcome of the replication is 0.41 (P=0.094, n=18). In Table S5 the correlations between market and survey beliefs and the other reproducibility indicators are also

shown, and Fig. S6 plots the relationship between beliefs and the relative effect size of the replications. The Spearman correlation coefficient between the market beliefs and the relative effect size is 0.28, but it is not significant (P=0.268, n=18). The Spearman correlation coefficient between the the pre-market survey beliefs and the relative effect size is 0.51 (P=0.030, n=18).

We also included a post-market survey to test if participating in the market affected the beliefs about reproducibility elicited in the survey. The post-market survey responses are very similar to the pre-market responses with a range of predictions from 57% to 83% and a mean of 70%. The Spearman correlation between the pre-market and the post-market survey is 0.96 (P<0.001, n=18), and the Spearman correlation between the post-market survey and the outcome of the replication is 0.58 (P=0.011, n=18). The absolute prediction error does not differ significantly between the pre-market survey and the post-market survey (Mean=0.418) (Wilcoxon signed-rank test, n=18, z=-1.55, P=0.122) or between the prediction market and the post market survey (Wilcoxon signed-rank test, n=18, z=-0.37, P=0.711).

The Spearman correlation between the pre-survey beliefs based on all the 140 individuals who filled out the survey and the 97 individuals who participated in the prediction markets is very high (0.99, P<0.001). The mean expected replication rate based on the survey beliefs for the sample of 140 individuals is 71% [range 54% to 87%, 95% CI =(66%, 76%)], and the Spearman correlation coefficient between the pre-market survey beliefs (n=140) and the outcome of the replication is 0.56 (P=0.016, n=18).

The relationship between the prediction market beliefs and survey beliefs and the replication outcomes can also be compared to a recent prediction market study on a subset of the studies (n=44) included in the RPP project (*26*). That study found a significant Pearson correlation between the prediction market beliefs and the replication outcomes of 0.42, compared to the non-significant Spearman correlation of 0.30 in this study (the Pearson correlation is 0.29 (P=0.247)). For the survey the Pearson correlation to the replication outcomes was 0.27 and non-significant in the RPP prediction markets, compared to the significant Spearman correlation of 0.52 in this study (the Pearson correlation is 0.49 (P=0.037)). Based on the point estimates the survey thus performs relatively better in this study compared to the RPP prediction markets.

But in comparing the results across the two prediction markets studies and in interpreting the non-significant positive association between the market beliefs and the replication outcomes in this study, it is important to bear in mind that the statistical power to find a significant correlation is limited in this sample due to the small sample size (n=18) and the relatively small variation in the prediction markets beliefs. To estimate this power we perform a simulation drawing 10,000 independent samples (n=18 in each draw) from our 18 observations where the actual replication probability for each study is exactly its prediction market belief and then we calculate the Spearman correlation coefficient and its p-value in each draw. With this method we estimate a power of around 15% to detect a significant correlation. The average correlation is around 0.25 in the 10,000 draws.

6. Comparison of reproducibility indicators between experimental economics and psychological sciences

We compared the results for the six reproducibility indicators included in the study to the results for psychological sciences in the RPP project (reported in Fig. 4). To test if the fraction of studies that "Replicated with P<0.05 in original direction" differed between the studies (61.1% (11/18) versus 36.1% (35/97)) we used a chi-square test that was significant (χ^2=3.96; P=0.047). To test if the "Original effect size within replication 95% CI" differed between the studies (66.7% (12/18) versus 47.4% (45/95)) we used the same test, but this was not significant (χ^2=2.25; P=0.133). The same test was also used to test if the "Meta-analytic estimate significant in the original direction" differed between the studies (77.8% (14/18) versus 68.0% (51/75)), but this difference was not significant (χ^2=0.66; P=0.417). The difference in "Replication effect-size (% of original effect size)" was compared using a Mann-Whitney U test, but this difference (65.7% (n=18) versus 44.5% (n=94)) was not significant (z=1.39, P=0.166). The same test was used to compare "Prediction market beliefs" between the studies, and this difference (75.1% (n=18) versus 55.1% (n=44)) was significant (z=3.21, P=0.001). Also for "Survey beliefs" the same test was used, and this difference (71.1% (n=18) versus 53.7% (n=43)) was also significant (z=4.89, P<0.001).

Note that one drawback of the reproducibility indicator "Original effect size within replication 95% CI" is that it does not include studies where the original estimate is below the 95% CI of the replication. We had one such replication that is counted as a successful replication with the indicator "Replicated with P<0.05 in original direction", but not with the indicator "Original effect size within replication 95% CI". But for comparability with the RPP results we still include the reproducibility indicator "Original effect size within replication 95% CI".

The results for the RPP study are taken from the published replication results (*19*) and the published prediction markets and survey results (*26*). The RPP project did not directly report the relative effect size of the replication, but instead used the "effect size difference" as a reproducibility indicator. The "effect size difference" was estimated as the absolute difference in the standardized effect size (r) between the original study and the replication study. We prefer to use the relative effect size (the ratio between the standardized effect size (r) of the replication and the standardized effect size (r) of the original study). The reason for this is the lack of comparability of the standardized effect sizes between our 18 studies discussed in section 2 above; caused by the difference in the level of aggregation of individual observations between the studies. To estimate the relative effect size from the RPP study we downloaded their posted effect size data and estimated the relative replication effect of each study. The original studies reporting null results in the RPP study (n=3) were excluded from this estimation; as we only included original results reporting positive results in our replication project.

7. Results and data for the individual studies/markets

Detailed replication results for the 18 studies are shown in Table S1. The hypotheses as described to the participants on the prediction markets in each of the 18 studies are shown in Table S2. In Table S3 we present the market belief, the statistical power of the replication, and the survey results for each of the 18 studies. Additional prediction market data are shown in Table S4. In Table S5 we also provide a correlation matrix for the six reproducibility indicators and the two original study characteristics included in the analyses.

To test the robustness of the correlations between the two original study characteristics (the p-value and the sample size) and the six reproducibility indicators we also estimated these correlations after sequentially excluding each study (n=17 in all these correlations); i.e. we run the correlations again after removing one of the 18 observations (studies) and we do this for all the 18 observations (i.e. 18 robustness tests of the correlations). In these robustness tests the Spearman correlation (p-values) between the original p-value and the reproducibility indicators ranged between: -0.70 – -0.52 (0.002–0.034) for "Replicated P<0.05", -0.36 – -0.16 (0.161–0.544) for "Original within 95% CI", -0.55 – -0.36 (0.021–0.153) for "Meta-estimate P<0.05", -0.64 – -0.48 (0.006–0.052) for "Relative effect size (r)", -0.79 – -0.68 (<0.001–0.003) for "Market belief", -0.90 – -0.86 (<0.001–<0.001) for "Survey belief". The Spearman correlation (p-values) between the sample size and the reproducibility indicators ranged between: 0.58–0.71 (0.001–0.015) for "Replicated P<0.05", 0.42–0.63 (0.006–0.091) for "Original within 95% CI", 0.63–0.74 (0.001–0.007) for "Meta-estimate P<0.05", 0.75–0.84 (<0.001–0.001) for "Relative effect size (r)", 0.34–0.58 (0.014–0.186) for "Market belief", 0.67–0.86 (<0.001–0.004) for "Survey belief".

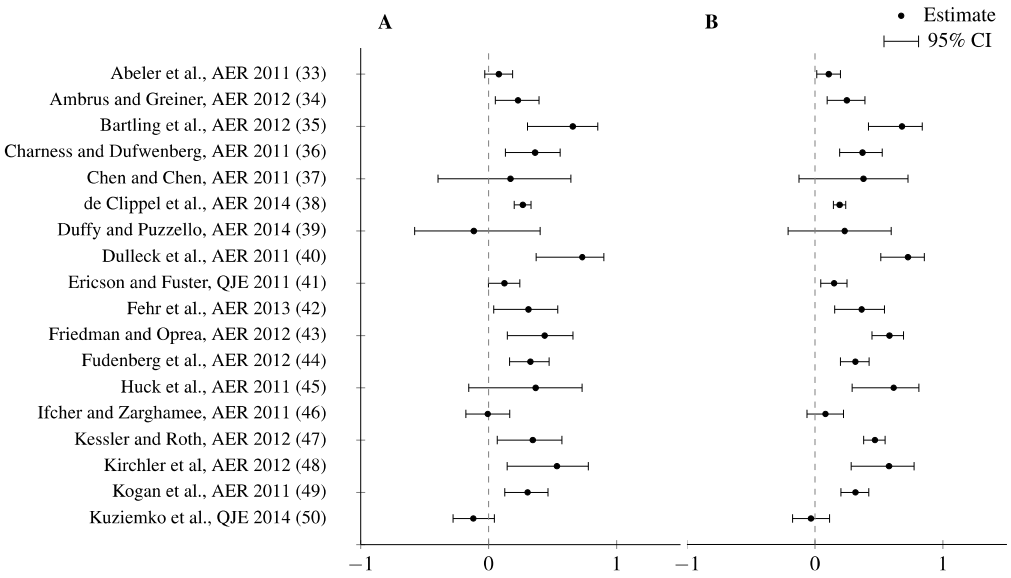

Fig. S1. A non-normalized version of Fig. 1 (Replication Results).
(**A**) 95% CIs of standardized replication effect sizes (correlation coefficient r).
(**B**) Meta-analytic estimates of effect sizes combining the original and replication studies. 95% CIs of standardized effect sizes (correlation coefficient r).

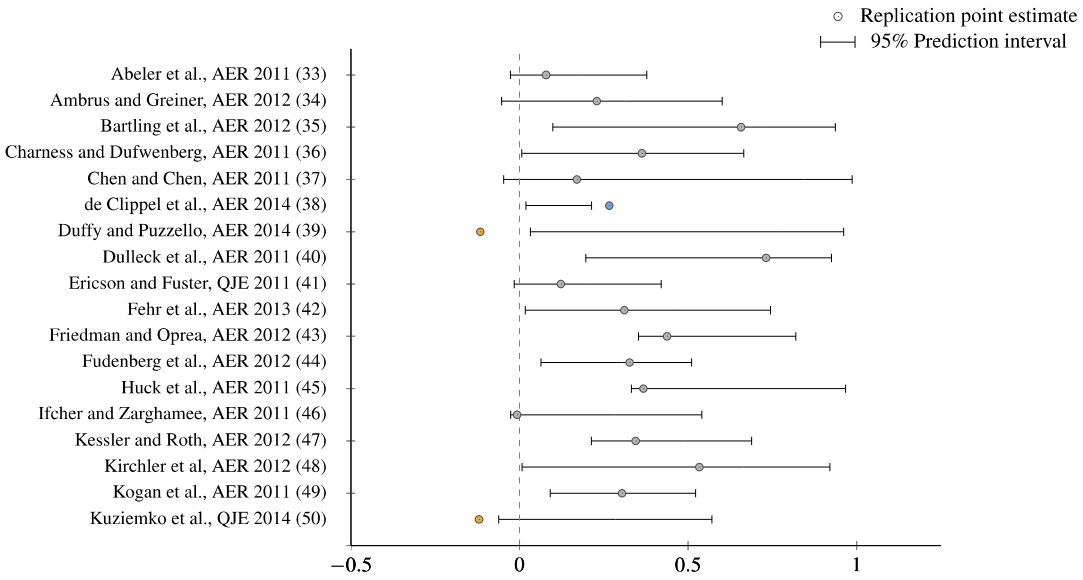

Fig. S2. 95% prediction intervals for the standardized original effect sizes (correlation coefficient r).

Fifteen replications (83.3%) are within the 95% prediction intervals; if we also include the replication with an effect size larger than the upper bound of the prediction interval this increases to 16 replications (88.9%).

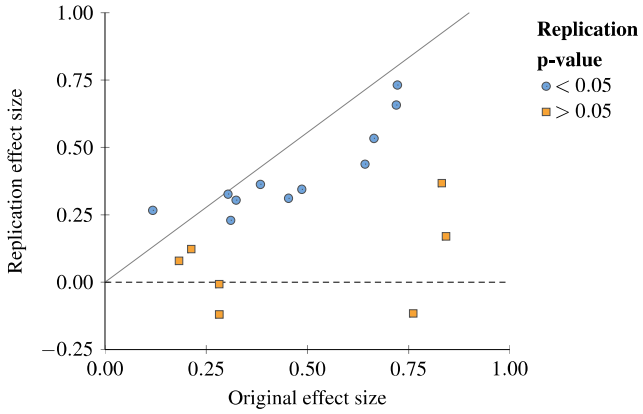

Fig. S3. Original study effect size versus replication effect size (correlation coefficients r).

The diagonal line represents replication effect size equal to the original effect size and the dotted line represents a replication effect size equal to zero. Blue dots are the replications that were significant with P<0.05 in the original direction, and red dots are the replications that were not significant. The mean standardized effect size (correlation coefficient, r) of the replications is 0.279 (SD=0.234), compared to 0.474 (SD=0.239) in the original studies. This difference is significant (Wilcoxon signed-ranks test, n=18, z=-2.98, P=0.003). The mean relative effect size of the replications is 65.9% [95% CI=(37.2%, 94.7%)]. The Spearman correlation between the original effect size and the replication effect size is 0.48 (P=0.043).

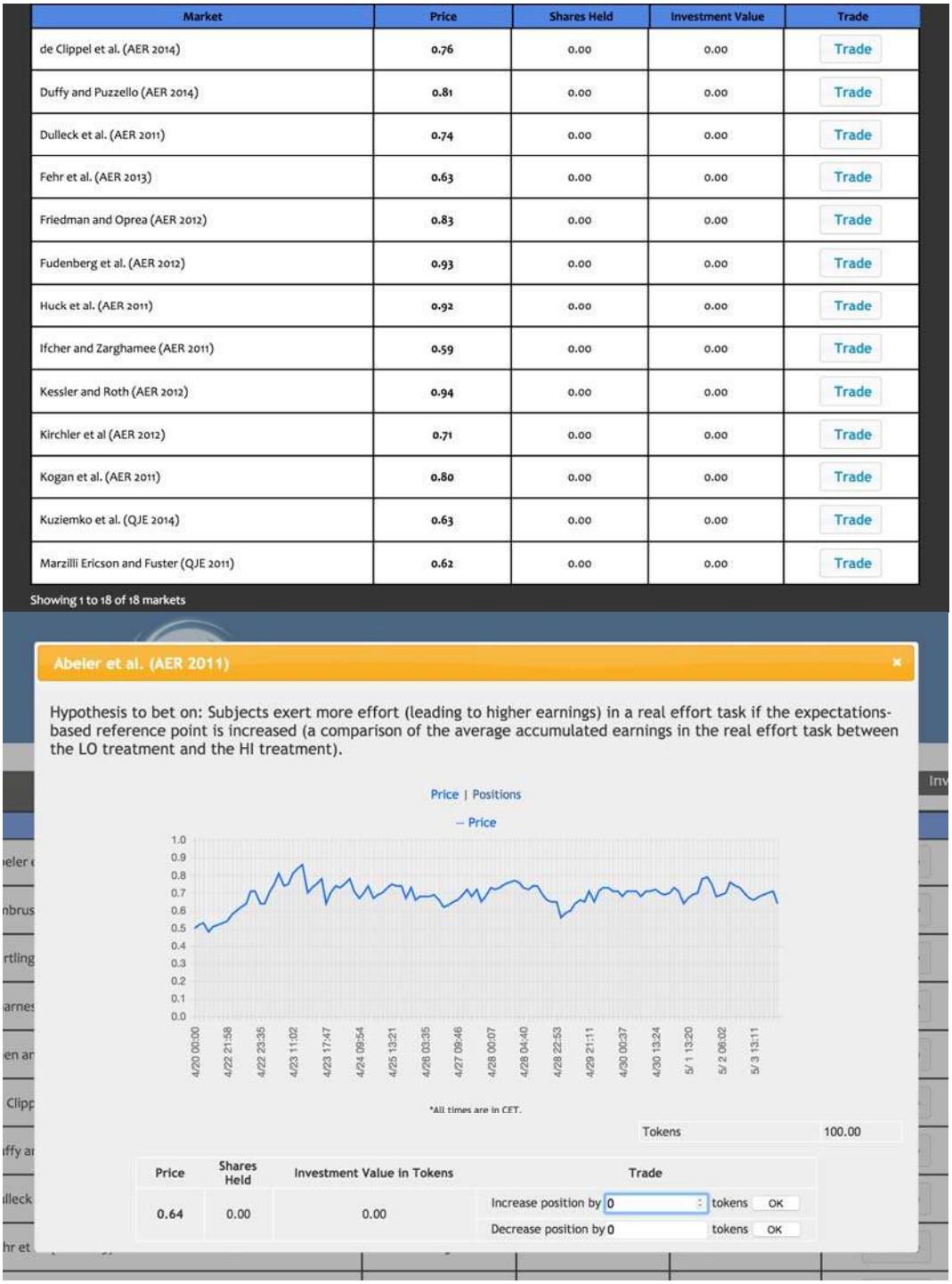

Fig. S4. Trading interface.
(**Top**) Market overview.
(**Bottom**) Trading page.

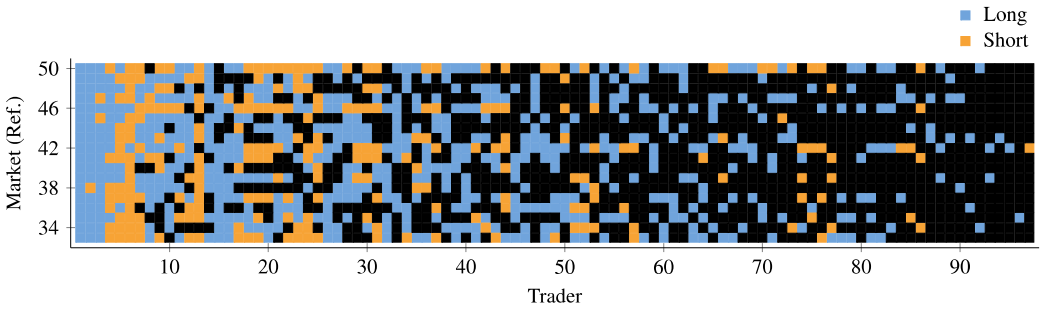

Fig. S5. Final positions per participant and market.

Long positions (bets on success) are shown in blue and short positions (bets on failure) are shown in red. The participants typically had broad portfolios with positions in several markets, and each market attracted a number of traders. The views differ between traders at the closing of the markets: in each market, there is at least one trader holding a long position, and one trader holding a short position. There are a few "bears" (predominantly betting on failure) who invested only in short positions (3/97 traders), and a larger fraction of "bulls" (predominantly betting on success) who invested only in long positions (40/97 traders). The majority of the participants fall into a wide spectrum between these two extremes.

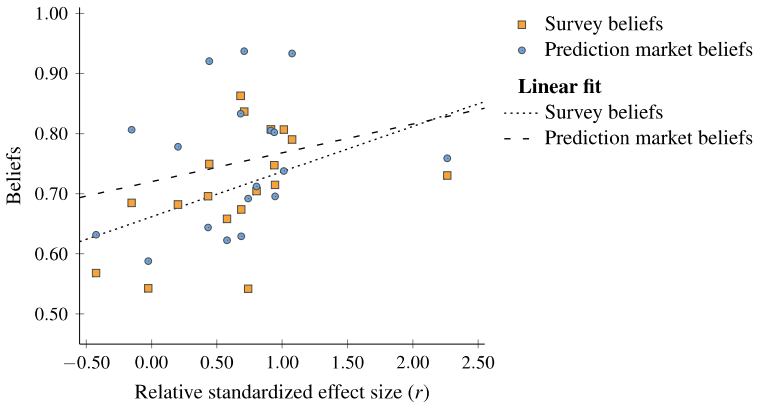

Fig. S6. Prediction market and survey beliefs and the relative effect size.

Both the prediction market beliefs (Spearman correlation coefficient 0.28, P=0.268, n=18), and the survey beliefs (Spearman Correlation Coefficient 0.51, P=0.030, n=18) are positively correlated with the relative effect size of the replications.

Table S1. Replication results.

		Original			Replication				Relative	
Study	**Ref.**	**P-value**	**Effect Size (*r*)**	**N***	**P-value**	**Effect Size (*r*)**	**N***	**Replicated**	**Effect Size (*r*)#**	
Abeler et al. (AER 2011)	33	0.046	0.18	120	0.160	0.08	318	No	0.43	(0.36)
Ambrus and Greiner (AER 2012)	34	0.057	0.31	117	0.012	0.23	357	Yes	0.74	(0.69)
Bartling et al. (AER 2012)	35	0.007	0.72	216	0.001	0.66	360	Yes	0.91	(1.21)
Charness and Dufwenberg (AER 2011)	36	0.010	0.38	162	0.003	0.36	264	Yes	0.95	(0.93)
Chen and Chen (AER 2011)	37	0.033	0.84	72	0.571	0.17	168	No	0.20	(0.22)
de Clippel et al. (AER 2014)	38	0.001	0.12	158	<0.001	0.27	156	Yes	2.26	(3.21)
Duffy and Puzzello (AER 2014)	39	0.010	0.76	54	0.674	−0.12	96	No	−0.15	(-0.19)
Dulleck et al. (AER 2011)	40	<0.001	0.72	168	0.001	0.73	128	Yes	1.01	(0.94)
Ericson and Fuster (QJE 2011)	41	0.030	0.21	112	0.055	0.12	262	No	0.58	(0.69)
Fehr et al. (AER 2013)	42	0.011	0.45	60	0.026	0.31	102	Yes	0.69	(0.84)
Friedman and Oprea (AER 2012)	43	<0.001	0.64	78	0.004	0.44	40	Yes	0.68	(0.68)
Fudenberg et al. (AER 2012)	44	0.001	0.30	124	<0.001	0.33	128	Yes	1.08	(0.96)
Huck et al. (AER 2011)	45	0.004	0.83	120	0.142	0.37	160	No	0.44	(0.43)
Ifcher and Zarghamee (AER 2011)	46	0.031	0.28	58	0.933	−0.01	131	No	−0.02	(-0.02)
Kessler and Roth (AER 2012)	47	<0.001	0.49	288	0.016	0.34	48	Yes	0.71	(0.62)
Kirchler et al (AER 2012)	48	0.016	0.66	120	0.010	0.53	220	Yes	0.80	(0.30)
Kogan et al. (AER 2011)	49	<0.001	0.32	126	0.001	0.30	90	Yes	0.94	(0.93)
Kuziemko et al. (QJE 2014)	50	0.070	0.28	42	0.154	−0.12	144	No	−0.42	(-0.39)

* N is the number of participants in the study. For the replications it is the actual rather than the planned number in the four replications where the actual sample size was somewhat higher than the planned sample size.

For completeness we report the relative non-standardized effect sizes in parenthesis. See the replication reports for more details.

Table S2. Hypotheses for the 18 replication studies.

Study	Ref.	Hypothesis
Abeler et al. (AER 2011)	33	Subjects exert more effort (leading to higher earnings) in a real effort task if the expectations-based reference point is increased (a comparison of the average accumulated earnings in the real effort task between the LO treatment and the HI treatment).
Ambrus and Greiner (AER 2012)	34	When there is imperfect monitoring, allowing punishment reduces net earnings (i.e., earnings after punishment costs).
Bartling et al. (AER 2012)	35	Adding a screening opportunity (informing the employer about the effort of the worker in the past three periods) for an employer that can offer full or limited discretion contracts increases efficiency (a comparison in average per-period total surplus between the base treatment and the screening treatment).
Charness and Dufwenberg (AER 2011)	36	Communication is effective in a hidden-information game when low-talent agents can participate in a Pareto-improving outcome (a comparison of the "Low B's Don't rate" for the messages (M) and no messages (NM) treatments for the (5,7) hidden information game).
Chen and Chen (AER 2011)	37	Effort in a minimum effort game is higher for subjects with a salient in-group identity than for subjects with a salient outgroup identity (a comparison of mean effort between the "Enhanced Ingroup" treatment and the "Enhanced Outgroup" treatment).
de Clippel et al. (AER 2014)	38	Efficiency (average aggregate payoff) is higher with the social choice mechanism Shortlisting (SL) than with the Veto-Rank (VR) mechanism for preference profile Pf2.
Duffy and Puzzello (AER 2014)	39	Efficiency in the Lagos-Wright money model is higher in an environment with money than in an environment without money for a population size of 6 (comparison in efficiency ratio between the money (M6) and the no money6 (NM6) treatments).
Dulleck et al. (AER 2011)	40	In a situation with verifiability, liability increases efficiency in a credence goods market (a comparison of efficiency between the B/LV treatment (liability/verifiability) and the B/V treatment (no liability/verifiability).
Ericson and Fuster (QJE 2011)	41	The willingness to accept (WTA) for a mug is higher for a high probability of receiving the mug for free compared to a low probability of receiving the mug for free (a comparison of the mean WTA between the treatment MH (80% chance of receiving the mug for free at the end of the experiment) and the treatment ML (10% chance of receiving the mug for free at the end of the experiment) in Experiment 2).
Fehr et al. (AER 2013)	42	The nonpecuniary disutility of being overruled causes a reluctance to delegate authority (a comparison of the average delegation rate between the HIGH NOREC and the PHIGH25 treatments).
Friedman and Oprea (AER 2012)	43	Cooperation in the prisoner's dilemma is higher in continuous time with flow payoffs over 60 seconds compared to eight equal subperiods (a comparison in the level of cooperation between the Continuous treatment and the Grid-8 treatment).
Fudenberg et al. (AER 2012)	44	Cooperation in a repeated prisoner's dilemma with noise (a specific probability that an intended move is changed to the opposite move) is higher when there are cooperative equilibria (a comparison in the level of overall cooperation between the b/c=1.5 and the b/c=2 treatment).
Huck et al. (AER 2011)	45	The ability to pay future deferred compensation increases worker earnings (w2+w3) more when commitment is enforced (FCT) compared to

		non-enforcement (NCT).
Ifcher and Zarghamee (AER 2011)	46	Showing subjects a film clip inducing positive affect will increase measured patience, excluding subjects who do not discount at all (Table 3, column 5).
Kessler and Roth (AER 2012)	47	An organ donation policy giving priority on waiting lists to those who previously registered as donors increase registered organ donors (a comparison of the fraction choosing to be a donor between the priority condition treatment and the control condition treatment in rounds 1-15 (the rounds for the between subjects comparison)).
Kirchler et al. (AER 2012)	48	A declining fundamental value (FV) increases mispricing in experimental asset markets (a comparison of the mean relative absolute deviation (RAD) between treatment 1 (T1) and treatment 2 (T2)).
Kogan et al. (AER 2011)	49	The presence of a preplay asset market lowers output in a "second-order statistic" coordination game (a comparison of group output of the insider groups in the market treatment and the control group in Experiment 2).
Kuziemko et al. (QJE 2014)	50	Subjects randomly placed in second-to-last place in terms of endowments are significantly less likely to allocate money to the person one rank below them in a choice of distributing $2 to the person one rank below or the person one rank above (a comparison of allocation decisions between subjects randomly ranked second-to-last and subjects randomly ranked 2-4 in the 6 person redistribution experiment).

Table S3. Prediction market and survey results for the 18 replication studies.

Study	Ref.	Replicated	Replication Power*		Market Belief	Pre-market Survey#		Post-market Survey
Abeler et al. (AER 2011)	33	No	0.90		0.644	0.696	(0.680)	0.697
Ambrus and Greiner (AER 2012)	34	Yes	0.90	(0.91)	0.692	0.542	(0.553)	0.620
Bartling et al. (AER 2012)	35	Yes	0.94		0.805	0.807	(0.800)	0.733
Charness and Dufwenberg (AER 2011)	36	Yes	0.90		0.695	0.715	(0.705)	0.708
Chen and Chen (AER 2011)	37	No	0.90		0.778	0.682	(0.674)	0.692
de Clippel et al. (AER 2014)	38	Yes	0.90		0.759	0.730	(0.730)	0.716
Duffy and Puzzello (AER 2014)	39	No	0.93		0.806	0.685	(0.683)	0.694
Dulleck et al. (AER 2011)	40	Yes	0.92		0.738	0.807	(0.811)	0.744
Ericson and Fuster (QJE 2011)	41	No	0.90	(0.91)	0.622	0.658	(0.649)	0.650
Fehr et al. (AER 2013)	42	Yes	0.91		0.629	0.674	(0.678)	0.666
Friedman and Oprea (AER 2012)	43	Yes	0.99		0.833	0.863	(0.866)	0.817
Fudenberg et al. (AER 2012)	44	Yes	0.90	(0.92)	0.933	0.790	(0.778)	0.770
Huck et al. (AER 2011)	45	No	0.91		0.920	0.749	(0.755)	0.730
Ifcher and Zarghamee (AER 2011)	46	No	0.90		0.588	0.542	(0.542)	0.566
Kessler and Roth (AER 2012)	47	Yes	0.95		0.937	0.837	(0.854)	0.825
Kirchler et al (AER 2012)	48	Yes	0.90		0.712	0.704	(0.712)	0.728
Kogan et al. (AER 2011)	49	Yes	0.94		0.802	0.748	(0.746)	0.752
Kuziemko et al. (QJE 2014)	50	No	0.91	(0.92)	0.632	0.568	(0.560)	0.582

* In a few studies the sample size in the replications was somewhat higher than the planned sample size. The planned statistical power is shown in this column, with the actual replication power in parentheses for those studies where it differed.

The average on the survey is shown for the 97 individuals who participated on the prediction market, and this is the variable used in the paper (unless stated otherwise). The value in parentheses is the average for all 140 individuals who filled in the survey. The Spearman correlation between the two pre-market survey variables is 0.99 ($P<0.001$).

Table S4. Additional prediction market data for the 18 replication studies.

Study	Ref.	Replicated	Market Belief	Volume	Tokens	Traders	Transactions
Abeler et al. (AER 2011)	33	No	0.644	1841	631	56	124
Ambrus and Greiner (AER 2012)	34	Yes	0.692	1515	486	48	101
Bartling et al. (AER 2012)	35	Yes	0.805	1401	559	50	95
Charness and Dufwenberg (AER 2011)	36	Yes	0.695	1180	394	37	84
Chen and Chen (AER 2011)	37	No	0.778	1826	584	52	119
de Clippel et al. (AER 2014)	38	Yes	0.759	735	255	31	81
Duffy and Puzzello (AER 2014)	39	No	0.806	1167	460	43	115
Dulleck et al. (AER 2011)	40	Yes	0.738	776	283	32	75
Ericson and Fuster (QJE 2011)	41	No	0.622	1787	458	47	122
Fehr et al. (AER 2013)	42	Yes	0.629	2849	946	67	193
Friedman and Oprea (AER 2012)	43	Yes	0.833	2247	688	49	188
Fudenberg et al. (AER 2012)	44	Yes	0.933	885	367	36	84
Huck et al. (AER 2011)	45	No	0.920	734	278	32	74
Ifcher and Zarghamee (AER 2011)	46	No	0.588	2371	735	56	165
Kessler and Roth (AER 2012)	47	Yes	0.937	1929	535	47	144
Kirchler et al (AER 2012)	48	Yes	0.712	866	281	37	78
Kogan et al. (AER 2011)	49	Yes	0.802	947	319	33	72
Kuziemko et al. (QJE 2014)	50	No	0.632	2684	867	68	166

Table S5. Correlation matrix for the six reproducibility indicators and the two original study characteristics included in the analyses.

Spearman correlations (P-values).

	Replicated P<0.05	Original within 95% CI	Meta-estimate P<0.05	Relative Effect Size (*r*)	Market Belief	Survey Belief	Original p-value	Original Sample Size
Replicated P<0.05	1.000							
Original within 95% CI	0.645	1.000						
	(0.004)							
Meta-estimate P<0.05	0.670	0.756	1.000					
	(0.002)	(0.000)						
Relative Effect Size (*r*)	0.846	0.522	0.721	1.000				
	(0.000)	(0.026)	(0.001)					
Market Belief	0.297	0.023	0.206	0.276	1.000			
	(0.232)	(0.929)	(0.412)	(0.268)				
Survey Belief	0.516	0.341	0.515	0.513	0.792	1.000		
	(0.028)	(0.166)	(0.029)	(0.030)	(0.000)			
Original p-value	-0.572	-0.273	-0.477	-0.561	-0.728	-0.879	1.000	
	(0.013)	(0.273)	(0.045)	(0.015)	(0.001)	(0.000)		
Original Sample Size	0.627	0.501	0.697	0.787	0.442	0.700	-0.611	1.000
	(0.005)	(0.034)	(0.001)	(0.000)	(0.067)	(0.001)	(0.007)	