Cite as: Camerer *et al.*, *Science* 10.1126/science.aaf0918 (2016).

Evaluating replicability of laboratory experiments in economics

Colin F. Camerer,[1*†] Anna Dreber,[2†] Eskil Forsell,[2†] Teck-Hua Ho,[3,4†] Jürgen Huber,[5†] Magnus Johannesson,[2†] Michael Kirchler,[5,6†] Johan Almenberg,[7] Adam Altmejd,[2] Taizan Chan,[8] Emma Heikensten,[2] Felix Holzmeister,[5] Taisuke Imai,[1] Siri Isaksson,[2] Gideon Nave,[1] Thomas Pfeiffer,[9,10] Michael Razen,[5] Hang Wu[4]

[1]California Institute of Technology, 1200 East California Boulevard, 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]Centre for Finance, Department of Economics, University of Göteborg, SE-40530 Göteborg, Sweden. [7]Sveriges Riksbank, SE-103 37 Stockholm, Sweden. [8]Office of the Deputy President (Research and Technology), National University of Singapore, Singapore 119077. [9]New Zealand Institute for Advanced Study, Private Bag 102904, North Shore Mail Centre, Auckland 0745, New Zealand. [10]Wissenschaftskolleg zu Berlin, Institute for Advanced Study, D-14193 Berlin, Germany.

***Corresponding author. E-mail: camerer@hss.caltech.edu**

†These authors contributed equally to this work.

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.

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 and 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 Fig. 2, fig. 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 Fig. 2 and fig. 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 0.05-.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). Medline doi:10.1126/science.1250475
2. J. P. A. Ioannidis, Why most published research findings are false. *PLOS Med.* **2**, e124 (2005). Medline doi:10.1371/journal.pmed.0020124
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 Discov.* **10**, 712 (2011). Medline doi:10.1038/nrd3439-c1
4. C. G. Begley, L. M. Ellis, Drug development: Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012). Medline doi:10.1038/483531a
5. L. P. Freedman, I. M. Cockburn, T. S. Simcoe, The economics of reproducibility in preclinical research. *PLOS Biol.* **13**, e1002165 (2015). Medline doi:10.1371/journal.pbio.1002165
6. K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, M. R. Munafò, Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013). Medline doi:10.1038/nrn3475
7. J. K. Hewitt, Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behav. Genet.* **42**, 1–2 (2012). Medline doi:10.1007/s10519-011-9504-z
8. M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, G. Getz, Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013). Medline doi:10.1038/nature12213
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**, 279–289 (1994). doi:10.1007/BF01175875
11. Z. Maniadis, F. Tufano, J. A. List, One swallow doesn't make a summer: New evidence on anchoring effects. *Am. Econ. Rev.* **104**, 277–290 (2014). doi:10.1257/aer.104.1.277
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**, 1359–1366 (2011). Medline doi:10.1177/0956797611417632
13. S. Carpenter, Psychology's bold initiative. *Science* **335**, 1558–1561 (2012). Medline doi:10.1126/science.335.6076.1558
14. Open Science Collaboration, An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. Psychol. Sci.* **7**, 657–660 (2012). Medline doi:10.1177/1745691612462588
15. J. Bohannon, Replication effort provokes praise—and 'bullying' charges. *Science* **344**, 788–789 (2014). Medline doi:10.1126/science.344.6186.788

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). Medline doi:10.1371/journal.pone.0029081
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). Medline doi:10.1371/journal.pone.0033423
18. B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, T. Yarkoni, Promoting an open research culture. *Science* **348**, 1422–1425 (2015). Medline doi:10.1126/science.aab2374
19. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015). Medline doi:10.1126/science.aac4716
20. A. Gelman, H. Stern, The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.* **60**, 328–331 (2006). doi:10.1198/000313006X152649
21. G. Cumming, Replication and *p* intervals: *P* values predict the future only vaguely, but confidence intervals do much better. *Perspect. Psychol. Sci.* **3**, 286–300 (2008). Medline doi:10.1111/j.1745-6924.2008.00079.x
22. J. Verhagen, E.-J. Wagenmakers, Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* **143**, 1457–1475 (2014). Medline doi:10.1037/a0036731
23. U. Simonsohn, Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* **26**, 559–569 (2015). Medline doi:10.1177/0956797614567341
24. J. T. Leek, P. Patil, R. D. Peng, http://arxiv.org/abs/1509.08968 (2015).
25. K. J. Arrow, R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. O. Ledyard, S. Levmore, R. Litan, P. Milgrom, F. D. Nelson, G. R. Neumann, M. Ottaviani, T. C. Schelling, R. J. Shiller, V. L. Smith, E. Snowberg, C. R. Sunstein, P. C. Tetlock, P. E. Tetlock, H. R. Varian, J. Wolfers, E. Zitzewitz, The promise of prediction markets. *Science* **320**, 877–878 (2008). Medline doi:10.1126/science.1157679
26. A. Dreber, T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, M. Johannesson, Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15343–15347 (2015). Medline doi:10.1073/pnas.1516179112
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**, 640–648 (2008). Medline doi:10.1097/EDE.0b013e31818131e7
29. B. D. McCullough, H. D. Vinod, Verifying the solution from a nonlinear solver: A case study. *Am. Econ. Rev.* **93**, 873–892 (2003). doi:10.1257/000282803322157133
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**, 1093–1107 (2006). doi:10.1353/mcb.2006.0061
32. A. Brodeur, M. Lé, M. Sangnier, Y. Zylberberg, Star Wars: The empirics strike back. *AEJ Applied* **8**, 1–32 (2016).
33. J. Abeler, A. Falk, L. Goette, D. Huffman, Reference points and effort provision. *Am. Econ. Rev.* **101**, 470–492 (2011). doi:10.1257/aer.101.2.470
34. A. Ambrus, B. Greiner, Imperfect public monitoring with costly punishment: An experimental study. *Am. Econ. Rev.* **102**, 3317–3332 (2012). doi:10.1257/aer.102.7.3317
35. B. Bartling, E. Fehr, K. M. Schmidt, Screening, competition, and job design: Economic origins of good jobs. *Am. Econ. Rev.* **102**, 834–864 (2012). doi:10.1257/aer.102.2.834
36. G. Charness, M. Dufwenberg, Participation. *Am. Econ. Rev.* **101**, 1211–1237 (2011). doi:10.1257/aer.101.4.1211
37. R. Chen, Y. Chen, The potential of social identity for equilibrium selection. *Am. Econ. Rev.* **101**, 2562–2589 (2011). doi:10.1257/aer.101.6.2562
38. G. de Clippel, K. Eliaz, B. G. Knight, On the selection of arbitrators. *Am. Econ. Rev.* **104**, 3434–3458 (2014). doi:10.1257/aer.104.11.3434
39. J. Duffy, D. Puzzello, Gift exchange versus monetary exchange: Theory and evidence. *Am. Econ. Rev.* **104**, 1735–1776 (2014). doi:10.1257/aer.104.6.1735
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**, 526–555 (2011). doi:10.1257/aer.101.2.526
41. K. M. Marzilli Ericson, A. Fuster, Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *Q. J. Econ.* **126**, 1879–1907 (2011). doi:10.1093/qje/qjr034
42. E. Fehr, H. Herz, T. Wilkening, The lure of authority: Motivation and incentive effects of power. *Am. Econ. Rev.* **103**, 1325–1359 (2013). doi:10.1257/aer.103.4.1325
43. D. Friedman, R. Oprea, A continuous dilemma. *Am. Econ. Rev.* **102**, 337–363 (2012). doi:10.1257/aer.102.1.337
44. D. Fudenberg, D. G. Rand, A. Dreber, Slow to anger and fast to forgive: Cooperation in an uncertain world. *Am. Econ. Rev.* **102**, 720–749 (2012). doi:10.1257/aer.102.2.720
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**, 819–843 (2011). doi:10.1257/aer.101.2.819
46. J. Ifcher, H. Zarghamee, Happiness and time preference: The effect of positive affect in a random-assignment experiment. *Am. Econ. Rev.* **101**, 3109–3129 (2011). doi:10.1257/aer.101.7.3109
47. J. B. Kessler, A. E. Roth, Organ allocation policy and the decision to donate. *Am. Econ. Rev.* **102**, 2018–2047 (2012). doi:10.1257/aer.102.5.2018
48. M. Kirchler, J. Huber, T. Stöckl, Thar she bursts: Reducing confusion reduces bubbles. *Am. Econ. Rev.* **102**, 865–883 (2012). doi:10.1257/aer.102.2.865
49. S. Kogan, A. M. Kwasnica, R. A. Weber, Coordination in the presence of asset markets. *Am. Econ. Rev.* **101**, 927–947 (2011). doi:10.1257/aer.101.2.927
50. I. Kuziemko, R. W. Buell, T. Reich, M. I. Norton, "Last-place aversion": Evidence and redistributive implications. *Q. J. Econ.* **129**, 105–149 (2014). doi:10.1093/qje/qjt035
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**, 793–827 (2012). doi:10.1093/qje/qjs013
52. C. C. Eckel, R. Petrie, Face value. *Am. Econ. Rev.* **101**, 1497–1513 (2011). doi:10.1257/aer.101.4.1497
53. D. Gill, V. Prowse, A structural analysis of disappointment aversion in a real effort competition. *Am. Econ. Rev.* **102**, 469–503 (2012). doi:10.1257/aer.102.1.469
54. N. Erkal, L. Gangadharan, N. Nikiforakis, Relative earnings and giving in a real-effort experiment. *Am. Econ. Rev.* **101**, 3330–3348 (2011). doi:10.1257/aer.101.7.3330
55. U. Fischbacher, z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* **10**, 171–178 (2007). doi:10.1007/s10683-006-9159-4
56. S. Palan, GIMS-Software for asset market experiments. *J. Behav. Exp. Finance* **5**, 1–14 (2015). Medline doi:10.1016/j.jbef.2015.02.001
57. R. Hanson, Could gambling save science? Encouraging an honest consensus. *Soc. Epistemology* **9**, 3–33 (1995). doi:10.1080/02691729508578768
58. J. Almenberg, K. Kittlitz, T. Pfeiffer, An experiment on prediction markets in science. *PLOS ONE* **4**, e8500 (2009). Medline doi:10.1371/journal.pone.0008500
59. J. Wolfers, E. Zitzewitz, Prediction markets. *J. Econ. Perspect.* **18**, 107–126 (2004). doi:10.1257/0895330041371321
60. G. Tziralis, I. Tatsiopoulos, Prediction markets: An extended literature review. *J. Pred. Mark.* **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. Mark.* **8**, 89 (2014).
63. C. F. Manski, Interpreting the predictions of prediction markets. *Econ. Lett.* **91**, 425–429 (2006). doi:10.1016/j.econlet.2006.01.004
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**, 11779–11784 (2013). Medline doi:10.1073/pnas.1206326110
65. R. Hanson, Logarithmic market scoring rules for modular combinatorial

information aggregation. *J. Pred. Mark.* **1**, 3 (2007).
66. Y. Chen, "Markets as an information aggregation mechanism for decision support," thesis, The Pennsylvania State University, State College, PA (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: Jorge Barraza, 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 S4 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. 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 analyzed 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.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/cgi/content/full/science.aaf0918/DC1
Materials and Methods
Figs. S1 to S6
Tables S1 to S5
References (*33*–*66*)

16 December 2015; accepted 19 February 2016
Published online 3 March 2016
10.1126/science.aaf0918

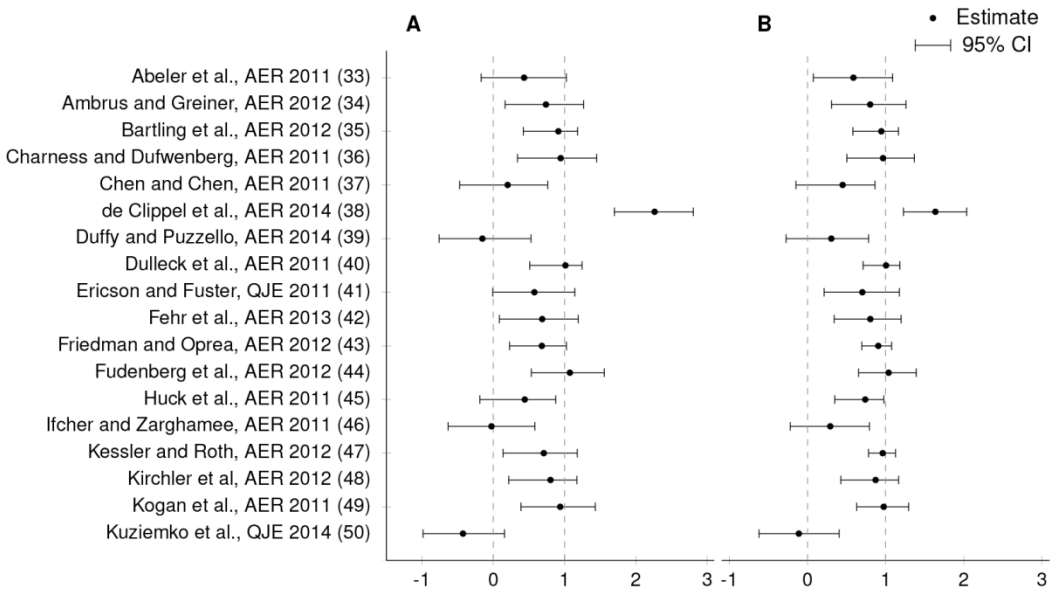

Fig. 1. Replication results. (**A**) Plotted are 95% CIs of replication effect sizes (standardized to correlation coefficients r). The standardized effect sizes are normalized so that 1 equals the original effect size (see fig. S1 for a non-normalized version). There is a significant effect in the same 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%)].

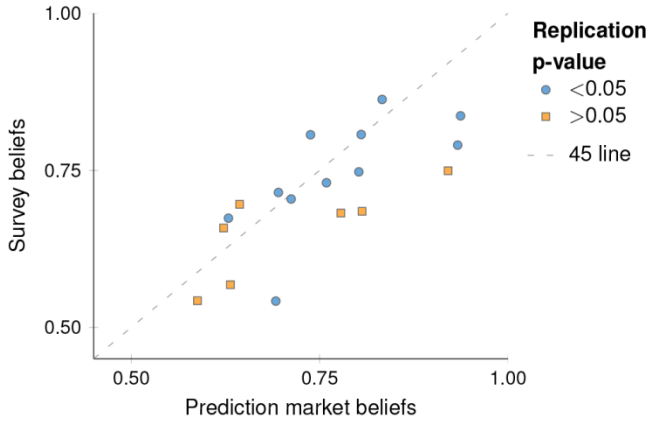

Fig. 2. Prediction market and survey beliefs. A plot of prediction market beliefs and survey beliefs in relation to if the original result was replicated with P<0.05 in the original direction. The mean prediction market belief is 75.2% [range 59% to 94%, 95% CI=(69.7%, 80.6%)], and the mean survey belief 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.

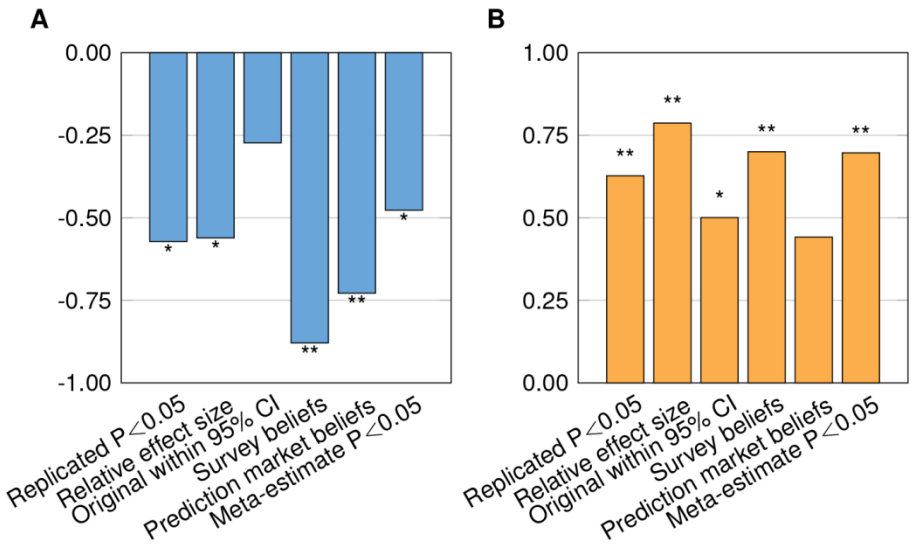

Fig. 3. Correlations between original study p-value and N and reproducibility indicators. The original p-value is negatively correlated with all six reproducibility indicators, and five of these correlations are significant. The original sample size is positively correlated with all six reproducibility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

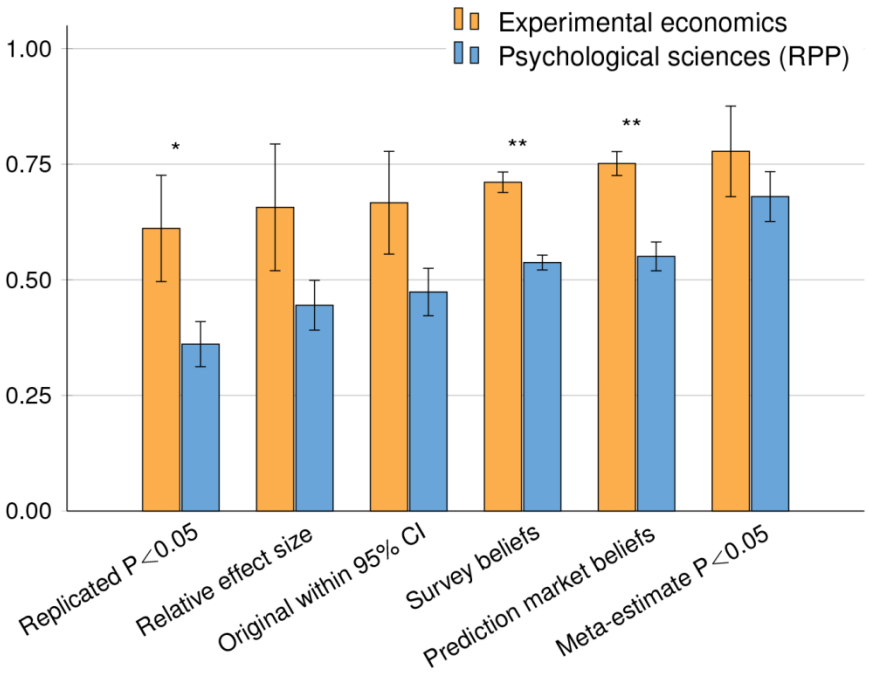

Fig. 4. A comparison of different reproducibility indicators between experimental economics and psychological sciences (the Reproducibility Project Psychology). Error bars denotes ±se. The reproducibility is higher for experimental economics for all six reproducibility indicators; this difference is 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.