**Do Economics Lab Experiments Replicate Predictably?**

*Laboratory experiments in Economics can be replicated predictably*

Colin F. Camerer1,\*,†, Anna Dreber2,†, Eskil Forsell2,†, Teck-Hua Ho3,4,†, Jürgen Huber5,†, Magnus Johannesson2,†, Michael Kirchler5,†, Johan Almenberg6, Adam Altmejd2, Emma Heikensten2, Felix Holzmeister5, Taisuke Imai1, Siri Isaksson2, Gidi Nave1, Thomas Pfeiffer7, Michael Razen5, Hang Wu4

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 Department of Economics, Faculty of Arts and Social Sciences, National University of Singapore, AS2#06-02, 1 Arts Link, Singapore 117570  5 Department of Banking and Finance, University of Innsbruck, Universitätsstrasse 15, 6020 Innsbruck, Austria 6 Sveriges Riksbank, SE-103 37 Stockholm, Sweden  
7 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

\* To whom correspondence should be addressed. E-mail: [camerer@hss.caltech.edu](mailto:camerer@hss.caltech.edu).

† These first seven authors contributed equally to this work.

**Abstract:** We replicate 18 laboratory experimental studies published in two high-impact influential economics journals in 2011-2014. All replications have a statistical power of ≥90% to detect the original effect size at the 5% significance level. We find a statistically significant effect in the same direction as in the original study for [1] replications, and on average the replicated effect size is [2]% of the original effect size. A prediction market is used to predict the outcome of the replications. The prediction market price is significantly correlated with the replication outcome and successfully predicts [3] of the 18 replications. The rate of reproducibility, both in terms of replication success and prediction market prices, is substantially higher than in a recent replication project in psychology.

**One Sentence Summary:** Results from 18 laboratory experimental studies published in high-impact economics journals were generally replicated ([4]% of the results) and peer predictions collected in pre-replication markets and surveys accurately forecasted when replication would be successful or not.

The deepest trust in scientific knowledge comes from the ability to directly replicate empirical findings, by using the same methods and getting the same results as others have. . While direct replication of this type is widely applauded [(McNutt 2014)](https://paperpile.com/c/tK6gyi/sh7a), it is rarely carried out in empirical social science. Replication is more important than ever, as the reproducibility of results has been questioned in many sciences, e.g. medicine [(Ioannidis 2005; Prinz, Schlange, and Asadullah 2011; Begley and Ellis 2012; Freedman, Cockburn, and Simcoe 2015)](https://paperpile.com/c/tK6gyi/6Guz+aX0i+zpRg+bGCd), neuroscience [(Button et al. 2013)](https://paperpile.com/c/tK6gyi/MrIK) and genetics [(Hewitt 2012; Lawrence et al. 2013)](https://paperpile.com/c/tK6gyi/peka+Nxaz).  Among social sciences, psychology has been the most active in both self-diagnosing the forces creating “false positives”, and conducting direct replications [(Simmons, Nelson, and Simonsohn 2011; Carpenter 2012; Open Science Collaboration 2012; Bohannon 2014)](https://paperpile.com/c/tK6gyi/Tgt8+5CGK+XEF4+QUtk). Several high-profile replication failures [(Doyen et al. 2012; Ritchie, Wiseman, and French 2012)](https://paperpile.com/c/tK6gyi/PKKt+jCoc) quickly led to changes in journal publication practices [(Nosek et al. 2015)](https://paperpile.com/c/tK6gyi/euwv). A recent Reproducibility Project Psychology (RPP) replicated 100 original studies published in three top journals in psychology. The majority (97)  of the replications reported “positive findings”; but only 35 (36%) of those were strong enough to be considered successful replications [(Open Science Collaboration 2015)](https://paperpile.com/c/tK6gyi/xWDn).

We provide substantial evidence about how well laboratory experiments in economics replicate. The sample consists ofis  18 experimental papers published in the American Economic Review and the Quarterly Journal of Economics in 2011-2014. We include all between- subject experimental comparisons published in these two leading journals in this time period. The most important statistically significant finding, as emphasized by the authors in each paper, is chosen for replication. We use replication sample sizes to have at least 90% power to detect the original effect size at the 5% significance level.

To assess whether experiments replicate reliably or not, it is also important to know whether peer scientists have accurate opinions about likelihood of replicability. To measure these opinions, before the 18 replications were done we conducted prediction markets before the 18 replications were done.   [(Wolfers and Zitzewitz 2004; Hanson 2007; Tziralis and Tatsiopoulos 2007; Arrow et al. 2008; Berg et al. 2008; Horn et al. 2014)](https://paperpile.com/c/tK6gyi/RAAV+CLXs+OXpN+BiCi+5sRb+DSRy). In the prediction market for a particular target study,  peers likely to be familiar with experimental methods in economics (see SOM for mailing details) could buy or sell artificial shares which pay money depending on whether the target study was replicated. The prediction markets produce a collective market probability of replication [(Manski 2006; Wolfers and Zitzewitz 2006)](https://paperpile.com/c/tK6gyi/bvTz+RCcR), and will provide evidence about how easy it is for peers to predict what papers will replicate or not [(Hanson 1995; Almenberg, Kittlitz, and Pfeiffer 2009; Dreber et al. 2015)](https://paperpile.com/c/tK6gyi/qPpL+s4Gt+LEzi). The traders’ survey beliefs about replicability were also collectedsurveyed, uncoupled from market trading  . Our studypaperThe project provides hard data on whether replication is successful, and whether successes and failures are predictable by peers.

The results of the replications are shown in Table 1. We find a statistically significant effect in the same direction as in the original study for [5] replications ([6]%; CI [7]-[8]).  The mean effect size of the replications is [9]% (SD=[10]) of the original effect size, with a variation between [11]% and [12]% as illustrated in Fig. 1. These results can be compared to the recently published results for psychological sciences, which found that 36% of “positive findings” in original studies were replicated with an average effect size of 49% of the original effect size [(Open Science Collaboration 2015)](https://paperpile.com/c/tK6gyi/xWDn). The fraction of studies that replicates in our study is significantly higher than the fraction for psychological sciences according to a contingency table chi-square test (chi-square value=[13]; P=[14]).

The Pearson correlation between the p-value of the original study and the P-value of the replication is [15] (P=[16]), and the sample size of the original study is also correlated with the P-value in the expected direction (Pearson correlation coefficient=[17], P=[18]).  These results are in line with those found for psychological sciences.

In the prediction markets, participants traded contracts that pay $1 if the study is replicated and $0 otherwise, and the resulting market price is interpreted as the predicted probability of the outcome occurring [(Manski 2006; Wolfers and Zitzewitz 2006)](https://paperpile.com/c/tK6gyi/bvTz+RCcR). 177 individuals originally signed up to participate, 140 of these filled in the pre-trading survey, 97 participated on the prediction market, and 80 participated in the post-trading survey. Each participant was endowed with USD 50 for trading. Further details about the prediction markets are given in the Supplementary Materials.

The mean prediction market final price is [19]% (range [20]% to [21]%), implying that about [22]% of the 18 original studies were expected to replicate (See Table S2 for more details). This is relatively close to the observed replication rate of [23]%, suggesting that the overall replication rate was captured well by the markets. When interpreting a market price larger than 50% as predicting successful replication and a market price smaller than 50% as predicting failed replication, we find that the prediction markets correctly predicts the outcome of [24]% of the replications ([25] of 18 studies; see Fig. 2), which is significantly higher than 50% (one-sample binomial test; P=[26]). The point-biserial correlation coefficient between the market price and the outcome of the replication is [27] and significant (P=[28], n=18).

To formally test if prediction market prices can be interpreted as probabilities of replication, we estimated a linear probability model (with robust standard errors) with the outcome of the replication as a function of the prediction market price. If market prices equal replication probabilities, the coefficient of the market price variable should be equal to one and the constant in the regression should be equal to zero. The coefficient of the market price variable is [29]; which is significantly different from zero (P=[30]), but not significantly different from 1 (P=[31]). The constant ([32]) is not significantly different from zero (t=[33], P=[34]).

The prediction market can also be compared to the pre-trading survey of participants’ beliefs about the probability of replication. A simple average of the survey correctly predicts [35]% of outcomes ([36] of 18; see Fig. 2), which is significantly different from 50% (one-sample binomial test; P=[37]). The absolute prediction error is significantly lower for the prediction market than for the survey (paired t-test, n=18, t=[38], P=[39]); see the Supplementary Materials for a more detailed comparison of the prediction market and survey responses).

The market prices can also be used to derive the probability of a tested hypothesis being true, also referred to as the positive predictive value [(Button et al. 2013)](https://paperpile.com/c/tK6gyi/MrIK), at different testing stages (see, [(Dreber et al. 2015)](https://paperpile.com/c/tK6gyi/LEzi)). Using information about the power and significance levels of the original study and the replications, the probability of a tested hypothesis being true is estimated for three stages of the testing process: the prior probability (p0) before observing the outcome of the initial study; the probability after observing the result of the initially published study (p1); and the probability after observing the outcome of the replication (p2). These results are shown in Fig. 3 and in Table S1.

The mean prior is [40]% with a range between [41]% and [42]%. After observing the positive finding in the first study the probability of the hypothesis being true (p1) increases to a mean of [43]% (range [44]% to [45]%). This estimate implies that about [46]% of statistically significant research findings published in top economics journals can be expected to be false positives. For the [47] original results that replicated the probability of the hypothesis being true (p2)increases further to a mean of [48]% (range [49]% to [50]%); and for the [51] original results that failed to be replicated the probability decreases to [52] (range [53]% to [54]%).

In line with the high rate of original results replicating in our sample compared to psychological sciences, the estimated priors derived from the prediction markets is also substantially higher in our sample. This suggests that top economics journals on average are less likely to publish surprising findings than top psychology journals. But there are some exceptions to this with the [55] studies in our economics sample having estimated priors below 25%, and out of these studies only [56] replicated.

**Discussion**

Another lesson is that replication is especially difficult to carry out when designs are not chosen, or not documented well enough, to make replication easy to do, perhaps a long time in the future.

These results provide evidence about two questions: (1) Do laboratory experiments in economics generally replicate? And (2) do peers accurately guess which studies will replicate or not?

The answer to question (1) is that Replication in these economics experiments is much better than in recent waves of psychology studies [(Open Science Collaboration 2012 plus more recent references)](https://paperpile.com/c/tK6gyi/XEF4). We think that this enhanced replicability could be because experimentation in  economics evolved much later than in psychology, in the 1970s, and developed special features. Economics experiments typically test a treatment effect that is clearly motivated by mature formal theory. Since such theories are taken seriously, a failure to find a predicted treatment effect can be an important null result; it can be published and not buried in the proverbial file drawer. Experimental economists also adopted strong norms about motivating subjects with substantial financial incentives, and treating deception as taboo, which probably enhance internal validity of treatment when different labs repeat the same experiments, either directly or conceptually. Pioneering experimental economists also persuaded journals to print instructions (and even raw data) in scarce journal pages, creating norms of transparency in sharing designs and data.

Just as optical illusions can be reproduced easily in psychology classes, simple economics experiments are reproduced in classes in order to illustrate principles. Without robust experimental methods, these classroom illustrations will backfire.

The answer to question (2) is that scientific peers *did* know which experiments were most likely to replicate (as expressed by market and surveyed beliefs). This evidence implies that some results published in widely-cited prestigious journals were known to be unlikely to replicate; and we found out which studies were the weak ones just by asking peers.

Evidence of accurate peer-based predictability suggests ways to improve pre-publication evaluation of scientific results.  Different types of peer evaluation conceivable use different judgments of how many, and which, peer opinions are included. The standard practice is for an editor to pick a small number of peers who are thought to be highly-informed, are expected to be objective, and produce thorough reports. Our results suggest that asking more peers for shorter ‘reports’ (even a single numerical judgment of replicability) could add useful information to the standard practice.

***BIG DRAMATIC FINISH***

\*leave survey versus PM measures for future work …

**References**

[Almenberg, Johan, Ken Kittlitz, and Thomas Pfeiffer. 2009. “An Experiment on Prediction Markets in Science.” *PloS One* 4 (12): e8500.](http://paperpile.com/b/tK6gyi/s4Gt)

[Arrow, Kenneth J., Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, and Forrest D. Nelson. 2008. “The Promise of Prediction Markets.” *SCIENCE-NEW YORK THEN WASHINGTON-* 320 (5878): 877.](http://paperpile.com/b/tK6gyi/BiCi)

[Begley, C. Glenn, and Lee M. Ellis. 2012. “Drug Development: Raise Standards for Preclinical Cancer Research.” *Nature* 483 (7391): 531–33.](http://paperpile.com/b/tK6gyi/zpRg)

[Berg, Joyce, Robert Forsythe, Forrest Nelson, and Thomas Rietz. 2008. “Results from a Dozen Years of Election Futures Markets Research.” *Handbook of Experimental Economics Results* 1: 742–51.](http://paperpile.com/b/tK6gyi/5sRb)

[Bohannon, John. 2014. “Replication Effort Provokes Praise—and ‘bullying’ Charges.” *Science* 344 (6186): 788–89.](http://paperpile.com/b/tK6gyi/QUtk)

[Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. “Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience.” *Nature Reviews. Neuroscience* 14 (5): 365–76.](http://paperpile.com/b/tK6gyi/MrIK)

[Carpenter, Siri. 2012. “Psychology’s Bold Initiative.” *Science* 335 (6076): 1558–61.](http://paperpile.com/b/tK6gyi/5CGK)

[Doyen, Stéphane, Olivier Klein, Cora-Lise Pichon, and Axel Cleeremans. 2012. “Behavioral Priming: It’s All in the Mind, but Whose Mind.” *PloS One* 7 (1): e29081.](http://paperpile.com/b/tK6gyi/PKKt)

[Dreber, Anna, Thomas Pfieffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. 2015. “Using Prediction Markets to Estimate the Reproducibility of Scientific Research.”](http://paperpile.com/b/tK6gyi/LEzi)

[Freedman, Leonard P., Iain M. Cockburn, and Timothy S. Simcoe. 2015. “The Economics of Reproducibility in Preclinical Research.” *PLoS Biology* 13 (6): e1002165.](http://paperpile.com/b/tK6gyi/bGCd)

[Hanson, Robin. 1995. “Could Gambling Save Science? Encouraging an Honest Consensus.” *Social Epistemology: A Journal of Knowledge, Culture and Policy* 9 (1): 3–33.](http://paperpile.com/b/tK6gyi/qPpL)

[———. 2007. “Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation.” *The Journal of Prediction Markets* 1 (1): 3–15.](http://paperpile.com/b/tK6gyi/CLXs)

[Hewitt, John K. 2012. “Editorial Policy on Candidate Gene Association and Candidate Gene-by-Environment Interaction Studies of Complex Traits.” *Behavior Genetics* 42 (1): 1–2.](http://paperpile.com/b/tK6gyi/peka)

[Horn, Christian Franz, Bjoern Sven Ivens, Michael Ohneberg, and Alexander Brem. 2014. “Prediction Markets – A Literature Review 2014.” *The Journal of Prediction Markets* 8 (2): 89–126.](http://paperpile.com/b/tK6gyi/DSRy)

[Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLoS Medicine* 2 (8): e124.](http://paperpile.com/b/tK6gyi/6Guz)

[Lawrence, Michael S., Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Chip Stewart, Craig H. Mermel, and Steven A. Roberts. 2013. “Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes.” *Nature* 499 (7457): 214–18.](http://paperpile.com/b/tK6gyi/Nxaz)

[Manski, Charles F. 2006. “Interpreting the Predictions of Prediction Markets.” *Economics Letters* 91 (3): 425–29.](http://paperpile.com/b/tK6gyi/bvTz)

[McNutt, Marcia. 2014. “Reproducibility.” *Science* 343 (6168): 229–229.](http://paperpile.com/b/tK6gyi/sh7a)

[Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, and G. Christensen. 2015. “Promoting an Open Research Culture: Author Guidelines for Journals Could Help to Promote Transparency, Openness, and Reproducibility.” *Science*  348 (6242): 1422.](http://paperpile.com/b/tK6gyi/euwv)

[Open Science Collaboration. 2012. “An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science.” *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 7 (6): 657–60.](http://paperpile.com/b/tK6gyi/XEF4)

[———. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251). doi:](http://paperpile.com/b/tK6gyi/xWDn)[10.1126/science.aac4716](http://dx.doi.org/10.1126/science.aac4716)[.](http://paperpile.com/b/tK6gyi/xWDn)

[Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. “Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?” *Nature Reviews. Drug Discovery* 10 (9): 712–712.](http://paperpile.com/b/tK6gyi/aX0i)

[Ritchie, Stuart J., Richard Wiseman, and Christopher C. French. 2012. “Failing the Future: Three Unsuccessful Attempts to Replicate Bem’s ‘retroactive Facilitation of Recall’effect.” *PloS One* 7 (3): e33423.](http://paperpile.com/b/tK6gyi/jCoc)

[Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science*, 0956797611417632.](http://paperpile.com/b/tK6gyi/Tgt8)

[Tziralis, Georgios, and Ilias Tatsiopoulos. 2007. “Prediction Markets: An Extended Literature Review.” *The Journal of Prediction Markets* 1 (1): 75–91.](http://paperpile.com/b/tK6gyi/OXpN)

[Wolfers, Justin, and Eric Zitzewitz. 2004. “Prediction Markets.” *The Journal of Economic Perspectives: A Journal of the American Economic Association* 18 (2): 107–26.](http://paperpile.com/b/tK6gyi/RAAV)

[———. 2006. “Interpreting Prediction Market Prices as Probabilities.” National Bureau of Economic Research.](http://paperpile.com/b/tK6gyi/RCcR)

**Additional SM refs:**

Merlob, B., Plott, C. R. & Zhang, Y. The CMS auction: experimental studies of a median-bid procurement auction with nonbinding bids. Quarterly Journal of Economics 127, 793-827 (2012).

Gill, D. & Prowse, V. A structural analysis of disappointment aversion in a real effort competition. American Economic Review 102, 469-503 (2012).

Eckel, C. C. & Petrie, R. Face value. American Economic Review 101, 1497-1513 (2011).

Erkal, N., Gangadharan, L. & Nikiforakis, N. Relative earnings and giving in a real-effort experiment. American Economic Review 101, 3330-3348 (2011).

Ifcher and Zarghamee, AER 2011;101(7):3109-3129.

Abeler., et al, AER 2011;101(2):470-492.

Dulleck et al., AER 2011;101(2):526-555.

Huck et al., AER 2011;101(2):819-843.

Kogan et al., AER 2011;101(2):927-947.

Charness and Dufwenberg., AER 2011;101(4);1211-1237.

Chen and Chen, AER 2011;101(6):2562-2589.

Ifcher and Zarghamee AER 2011;101(7):3109-3129.

Friedman and Oprea, AER 2012;102(1):337-363.

Fudenberg et al., AER 2012;102(2):720-729.

Bartling et al., AER 2012;102(2):834-864.

Kirchler et al, AER 2012;102(2):865-883.

Kessler and Roth, AER 2012;102(5):2018-2047.

Ambrus and Greiner, AER 2012;102(7):3317-3332.

Fehr et al, AER 2013;103(4):1325-1359.

Duffy and Puzzello, AER 2014;104(6):1735-1776.

de Clippel et al, AER 2014;104(11):3434-3458.

Marzilli Ericson and Fuster, QJE 2011;126(4):1879-1907.

Kuziemko et al., QJE 2014;129(1):105-149.

**Acknowledgements:** We thank the Sloan Foundation (XX), Jan Wallander and Tom Hedelius Foundation (P2015-0001:1 and [P2013-0156:1](http://www.forskningsstiftelserna.3ddata.com/ansokan.asp?v1=4358)), the Knut and Alice Wallenberg Foundation (Wallenberg Academy Fellows grant to A. Dreber), the Swedish Foundation for Humanities and Social Sciences (NHS14-1719:1), the Behavioral and Neuroeconomics Discovery Fund (CC), the Austrian Science Fund FWF (START-grant Y617-G11) , etc… for financial support. The data reported in this paper are tabulated in Tables S1-S3 and the Replication Reports are available at XX and the data from the replications are available at XX. 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, TP helped design the prediction market part; EF, FH, JH, MK, MR, TP, XX analysed data; AA, EH, FH, 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](http://www.sciencemag.org)

Materials and Methods

Figs. S1-S4

Tables S1-S4

References: XX-XX

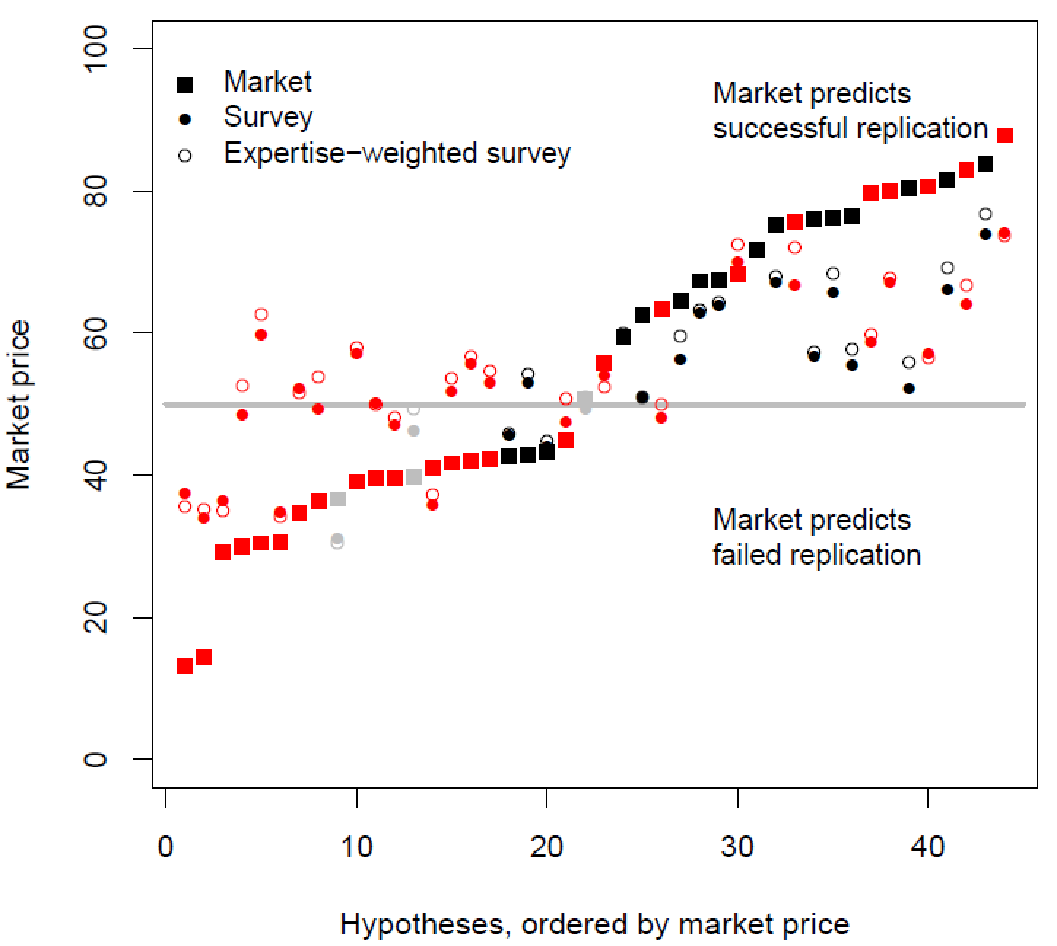
**TABLES**

**Table 1.** Replication results (hypothesis, original effect size, original p-value, replication effect size, replication p-value, relative effect size, replicated (yes/no)).

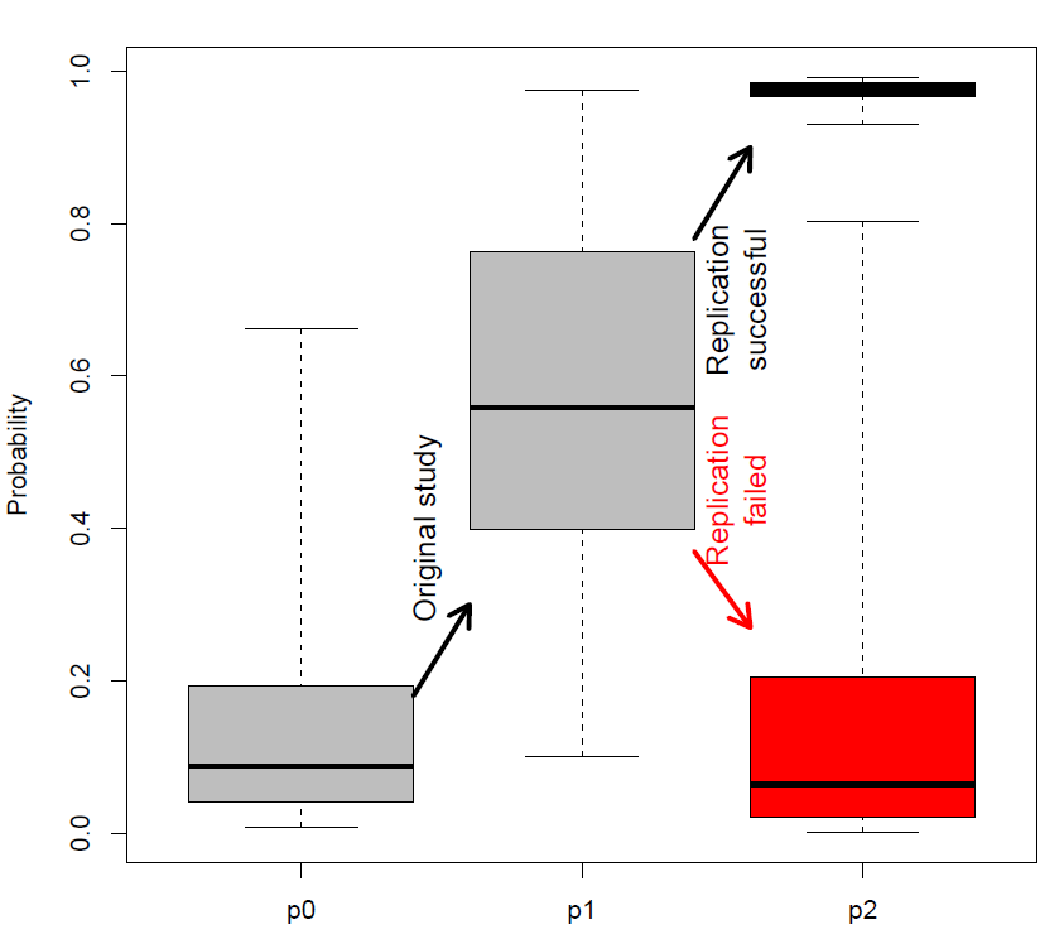
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Study  (ref number) | Original effect size | Replication effect size | Relative effect size | Original  p-value | Replication p-value | Replicated (yes/no) |
| Abeler et al. AER 2011(XX) |  |  |  | 0.046 |  |  |
| Dulleck et al. AER 2011(XX) |  |  |  | 0.0001 |  |  |
| Huck et al. AER 2011(XX) |  |  |  | 0.0039 |  |  |
| Kogan et al. AER 2011(XX) |  |  |  | 0.000026 |  |  |
| Charness et al. AER 2011(XX) |  |  |  | 0.010 |  |  |
| Chen & Chen AER 2011(XX) |  |  |  | 0.033 |  |  |
| Ifcher et al. AER 2011(XX) |  |  |  | 0.031 |  |  |
| Friedman et al. AER 2012(XX) |  |  |  | 4.0e-11 |  |  |
| Fudenberg et al. AER 2012(XX) |  |  |  | 0.001 |  |  |
| Bartling et al. AER 2012(XX) |  |  |  | 0.007 |  |  |
| Kirchler et al. AER 2012(XX) |  |  |  | 0.016 |  |  |
| Kessler & Roth AER 2012(XX) |  |  |  | 1.6e-18 |  |  |
| Ambrus et al. AER 2012(XX) |  |  |  | 0.057 |  |  |
| Fehr et al.  AER 2013(XX) |  |  |  | 0.011 |  |  |
| Duffy et al. AER 2014(XX) |  |  |  | 0.01 |  |  |
| de Clippel et al. AER 2014(XX) |  |  |  | 0.001 |  |  |
| Marzilli Ericson QJE 2011(XX) |  |  |  | 0.03 |  |  |
| Kuziemko et al. QJE 2014(XX) |  |  |  | 0.070 |  |  |

**FIGURES**

**Fig 1.** Relative effect size of replications (study number on X-axis and relative effect size on Y-axis; different colors for significant or non-significant or show exact p-value in graph).



**Fig. 2.** Prediction market performance. Final market prices and survey predictions are shown for the replication of 18 publications in American Economic Review and Quarterly Journal of Economics. The prediction market predicts XX out of 18 replications correctly. Successful replications (XX/18 replications) are shown in black and failed replications (XX/18) in red.



**Fig. 3.** Probability of a hypothesis being true at three different stages of testing: before the initial study (p0), after the initial study but before the replication (p1), and after replication (p2). “Error bars” (or whiskers) represent range, boxes are first to third quartiles, and thick lines are medians.