Supplementary Online Material

How Common are False Positives in Laboratory Economics Experiments? Evidence from the *P*-Curve Method

Taisuke Imai Klavdia Zemlianova Nikhil Kotecha
Colin F. Camerer

Contents

A The P-Curve Method: Implementation Detail

Simonsohn et al. (2014) propose to test for skewness of the observed p-curve to examine whether a set of studies contain evidential value.[1] The method takes variations of p-values within each bin into consideration by treating each individual p-value as a test statistic. We first compute, for each significant p-value, the probability of observing a significant p-value at least as extreme if the null hypothesis H_0, the flat p-curve, were true. It is called a pp-value, indicating the p-value of a p-value, and is computed simply as $pp = p/0.05$.[2] We then aggregate a set of pp-values from K studies, $(pp_k)_{k=1}^K$ into a single test statistic using Stouffer et al.'s (1949) method:

$$Z_S = \frac{\sum_{k=1}^K Z_k}{\sqrt{K}},$$

where $Z_k = \Phi^{-1}(1 - pp_k)$ where Φ is the standard normal cumulative distribution function. [3] Under the null hypothesis of a flat p-curve, the Stouffer test statistic Z_S follows the standard normal distribution $N(0, 1)$.

When a p-curve is not significantly right-skewed, one needs to distinguish following two possibilities: (i) studies under examination lack evidential value, or (ii) there is not enough information to make an inference about evidential value. In order to distinguish these two potential accounts, Simonsohn et al. (2014) proposed to test whether a p-curve is flatter than the one we would observe if studies have 33% power. Remember that a statistical power is the probability of a statistical test obtaining a p-value of less than α (the level or "size" of the test, typically set to 0.05) when the null is false:

$$\text{power} = \Pr[\text{reject } H_0 \mid H_1 \text{ is true}].$$

The cutoff power of 33% is chosen arbitrarily here: the idea is that we would be able to conclude that a set of studies lack evidential value if a p-curve is significantly flatter than the one we would expect to observe from a set of low-powered studies.

The actual implementation of the test directly follows the steps we described above. The only necessary twist is to compute pp-values under the null that a test has 33% power. Suppose we

[1] A simpler method mentioned by Simonsohn et al. (2014) tests for right-skewness is to split the set of significant p-values (i.e., $p \leq 0.05$) into a high ($p > 0.025$) bin and a low ($p \leq 0.025$) bin and then to apply the binomial test with the uniform null that half of the significant p-values falls into the high bin.

[2] For example, under the null hypothesis of flat p-curve, there is a 20% chance of observing $p < 0.01$.

[3] In Simonsohn et al. (2014), pp-values are aggregated using the Fisher's method , which follows a χ^2-distribution with $2K$ degree of freedom. Stouffer's method is used in the background computation for the p-curve online application (available at `http://www.p-curve.com`).

observe a test statistic g_k (with p-value p_k) in study k and the distribution of the test statistic is given by $G_{\mathbf{df}_k}$ (it can be F or χ^2 distribution). Here $\mathbf{df}_k$ is (potentially a vector of) degree of freedom(s) of the test for study k. We first identify the critical value $C_{0.05}$ for the test with level 0.05 such that $G_{\mathbf{df}_k}(C_{0.05}) = 1-0.05$. We then find the *noncentrality parameter* $\mathbf{ncp}_k$ for the distribution $G_{\mathbf{df}_k}$ that has a 33% power: the parameter that gives $G_{\mathbf{df}_k}(c \geq C_{0.05} \mid \mathbf{ncp}_k) = 1/3$.[4][5] We then evaluate the observed test statistic g_k with the derived noncentral distribution $G_{\mathbf{df}_k}(\cdot \mid \mathbf{ncp}_k)$, which gives the probability of observing a test statistic at most as extreme as the one we have (and hence the probability of having p-values larger than p_k) under the null of 33% power. Since the study is assumed to have a 33% power, there is a 2/3 chance of obtaining $p > 0.05$. Therefore, the desired pp-value for study k is computed by

$$pp_k = \frac{G_{\mathbf{df}_k}(g_k \mid \mathbf{ncp}_k) - 2/3}{1 - 2/3}.$$

See the supplementary materials for Simonsohn et al. (2014) for more details and concrete examples.[6]

[4]Noncentrality parameters connect central families and noncentral families of distributions. As is commonly used in statistical hypothesis testing, a central family is a distribution of a test statistic when the null hypothesis is true. A noncentral family, on the other hand, is a distribution of a test statistic when the null hypothesis is false. It is now clear from this definition that the noncentral distributions are closely related to the calculation of statistical power of the test.

[5]Note that the noncentrality parameter is indexed by study k, since it depends on the sample size of the study.

[6]Available here: `http://p-curve.com/`.

B Additional Results

Table B.1: Number of published papers by journal and year.

Journal	2009	2010	2011	2012	2013	2014	2015	2016
AEJ:Mic	2	1	2	2	1	1	0	1
AER	3	4	3	5	2	1	1	1
ECMA	1	0	0	0	0	1	1	0
EE	5	6	9	6	5	9	3	15
JPE	0	0	0	0	0	0	2	0
QJE	1	0	1	0	0	3	1	1
REStud	2	0	0	1	1	0	0	1

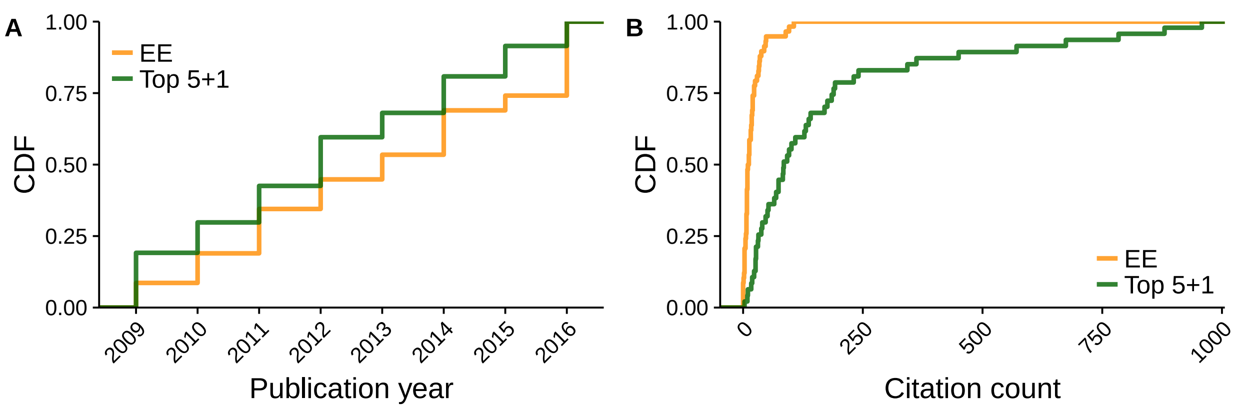

Figure B.1: (A) Cumulative distributions of the number of papers published in each year. (B) Cumulative distributions citation counts by journal category.

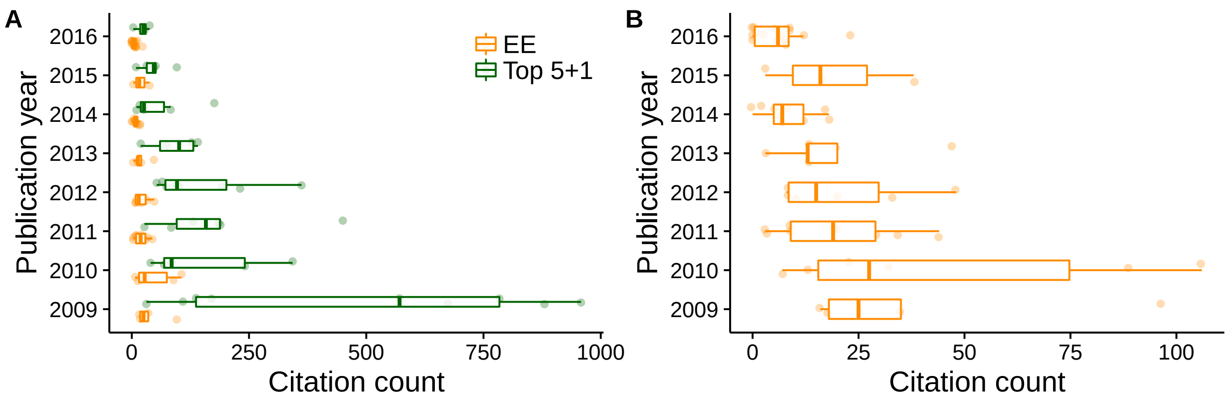

FIGURE B.2: Citation counts by year and journal. (A) Comparing EE and Top 5+1. (B) EE only.

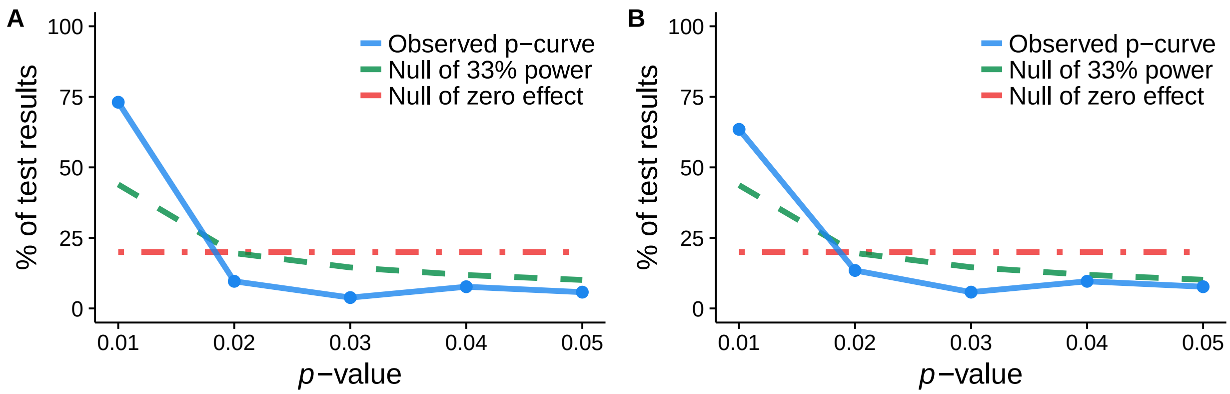

FIGURE B.3: *P*-curve: (A) below median citation counts vs. (B) above median citation counts.

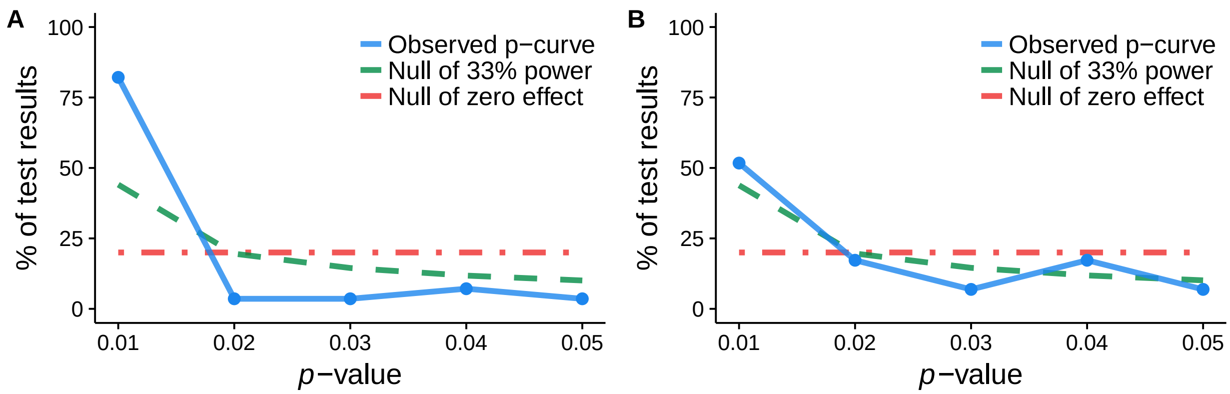

FIGURE B.4: *P*-curve from EE: (A) below median citation counts vs. (B) above median citation counts.

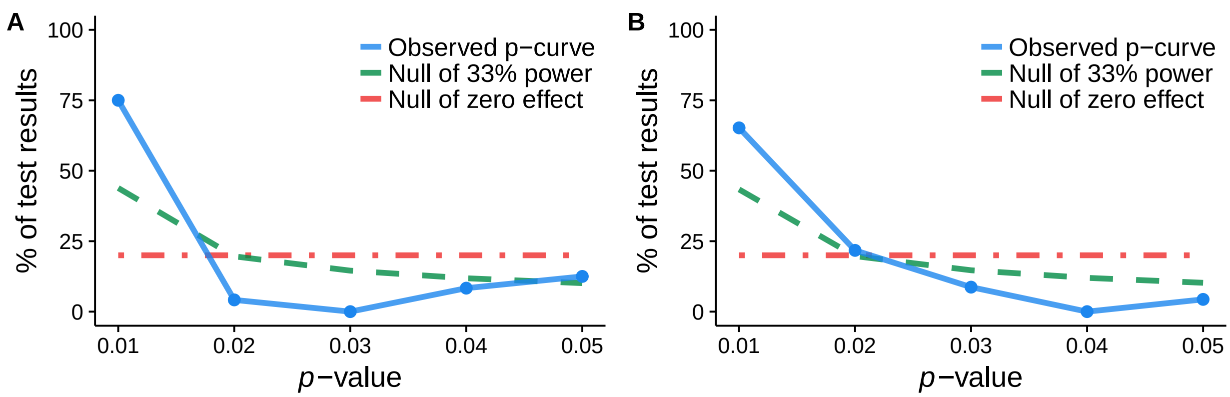

FIGURE B.5: *P*-curve from Top 5+1: (A) below median citation counts and (B) above median citation counts.

C List of Included Papers

Abeler, J., A. Falk, L. Goette, and D. Huffman (2011): "Reference Points and Effort Provision," *American Economic Review*, 101, 470–92.

Ackert, L. F., A. B. Gillette, J. Martinez-Vazquez, and M. Rider (2011): "Are Benevolent Dictators Altruistic in Groups? a Within-Subject Design," *Experimental Economics*, 14, 307–321.

Anderson, L. R., F. J. DiTraglia, and J. R. Gerlach (2011): "Measuring Altruism in a Public Goods Experiment: a Comparison of US and Czech Subjects," *Experimental Economics*, 14, 426–437.

Andreoni, J. and B. D. Bernheim (2009): "Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica*, 77, 1607–1636.

Andreoni, J. and C. Sprenger (2012): "Risk Preferences are Not Time Preferences," *American Economic Review*, 102, 3357–3376.

Arad, A. and A. Rubinstein (2012): "The 11–20 Money Request Game: A Level-*k* Reasoning Study," *American Economic Review*, 102, 3561–3573.

Ariely, D., A. Bracha, and S. Meier (2009a): "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, 99, 544–555.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar (2009b): "Large Stakes and Big Mistakes," *Review of Economic Studies*, 76, 451–469.

Asparouhova, E., P. Bossaerts, J. Eguia, and W. Zame (2015): "Asset Pricing and Asymmetric Reasoning," *Journal of Political Economy*, 123, 66–122.

Ballinger, T. P., E. Hudson, L. Karkoviata, and N. T. Wilcox (2011): "Saving Behavior and Cognitive Abilities," *Experimental Economics*, 14, 349–374.

Banerjee, R. (2016): "On the Interpretation of Bribery in a Laboratory Corruption Game: Moral Frames and Social Norms," *Experimental Economics*, 19, 240–267.

Bartling, B. and U. Fischbacher (2011): "Shifting the Blame: On Delegation and Responsibility," *Review of Economic Studies*, 79, 67–87.

Bartling, B., R. A. Weber, and L. Yao (2014): "Do Markets Erode Social Responsibility?" *Quarterly Journal of Economics*, 130, 219–266.

Becker, A. and L. M. Miller (2009): "Promoting Justice by Treating People Unequally: An Experimental Study," *Experimental Economics*, 12, 437–449.

Bellemare, C., L. Bissonnette, and S. Kröger (2010): "Bounding Preference Parameters under Different Assumptions About Beliefs: a Partial Identification Approach," *Experimental Economics*, 13, 334–345.

Benjamin, D. J., J. J. Choi, and A. J. Strickland (2010): "Social Identity and Preferences," *American Economic Review*, 100, 1913–1928.

Bigoni, M., M. Casari, A. Skrzypacz, and G. Spagnolo (2015): "Time Horizon and Cooperation in Continuous Time," *Econometrica*, 83, 587–616.

Bo, P. D., A. Foster, and L. Putterman (2010): "Institutions and Behavior: Experimental Evidence on the Effects of Democracy," *American Economic Review*, 100, 2205–2229.

Bolton, G. and P. Werner (2016): "The Influence of Potential on Wages and Effort," *Experimental Economics*, 19, 535–561.

Boly, A. (2011): "On the Incentive Effects of Monitoring: Evidence from the Lab and the Field," *Experimental Economics*, 14, 241–253.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016): "Stereotypes," *Quarterly Journal of Economics*, 131, 1753–1794.

Botelho, A., A. Dinar, L. M. C. Pinto, and A. Rapoport (2014): "Time and Uncertainty in Resource Dilemmas: Equilibrium Solutions and Experimental Results," *Experimental Economics*, 17, 649–672.

Brandts, J., C. Rott, and C. Solà (2016): "Not Just Like Starting over-Leadership and Revivification of Cooperation in Groups," *Experimental Economics*, 19, 792–818.

Brown, A. L., Z. E. Chua, and C. F. Camerer (2009): "Learning and Visceral Temptation in Dynamic Saving Experiments," *Quarterly Journal of Economics*, 124, 197–231.

Brunner, C., J. K. Goeree, C. A. Holt, and J. O. Ledyard (2010): "An Experimental Test of Flexible Combinatorial Spectrum Auction Formats," *American Economic Journal: Microeconomics*, 2, 39–57.

Burks, S. V., J. P. Carpenter, L. Goette, and A. Rustichini (2013): "Overconfidence and Social Signalling," *Review of Economic Studies*, 80, 949–983.

Buser, T., M. Niederle, and H. Oosterbeek (2014): "Gender, Competitiveness, and Career Choices," *Quarterly Journal of Economics*, 129, 1409–1447.

Buser, T. and N. Peter (2012): "Multitasking," *Experimental Economics*, 15, 641–655.

Cabrales, A., G. Charness, and M. C. Villeval (2011): "Hidden Information, Bargaining Power, and Efficiency: An Experiment," *Experimental Economics*, 14, 133–159.

Cabrales, A., R. Nagel, and J. V. R. Mora (2012): "It is Hobbes, not Rousseau: An Experiment on Voting and Redistribution," *Experimental Economics*, 15, 278–308.

Camera, G. and M. Casari (2014): "The Coordination Value of Monetary Exchange: Experimental Evidence," *American Economic Journal: Microeconomics*, 6, 290–314.

Cappelen, A. W., J. Konow, E. Ø. Sørensen, and B. Tungodden (2013): "Just Luck: An Experimental Study of Risk-Taking and Fairness," *American Economic Review*, 103, 1398–1413.

Carpenter, J. and P. H. Matthews (2009): "What Norms Trigger Punishment?" *Experimental Economics*, 12, 272–288.

Carrillo, J. D. and T. R. Palfrey (2009): "The Compromise Game: Two-Sided Adverse Selection in the Laboratory," *American Economic Journal: Microeconomics*, 1, 151–181.

Charness, G., R. Cobo-Reyes, N. Jiménez, J. A. Lacomba, and F. Lagos (2012): "The Hidden Advantage of Delegation: Pareto Improvements in a Gift Exchange Game," *American Economic Review*, 102, 2358–2379.

Charness, G., R. Cobo-Reyes, J. A. Lacomba, F. Lagos, and J. M. Pérez (2016): "Social Comparisons in Wage Delegation: Experimental Evidence," *Experimental Economics*, 19, 433–459.

Charness, G. and M. Dufwenberg (2011): "Participation," *American Economic Review*, 101, 1211–1237.

Charness, G. and D. Levin (2009): "The Origin of the Winner's Curse: A Laboratory Study," *American Economic Journal: Microeconomics*, 1, 207–236.

Chen, Y., G. Y. Jeon, and Y.-M. Kim (2014): "A Day without a Search Engine: An Experimental Study of Online and Offline Searches," *Experimental Economics*, 17, 512–536.

Chen, Y. and S. X. Li (2009): "Group Identity and Social Preferences," *American Economic Review*, 99, 431–457.

Cheung, S. L. (2014): "New Insights into Conditional Cooperation and Punishment from a Strategy Method Experiment," *Experimental Economics*, 17, 129–153.

Cleave, B. L., N. Nikiforakis, and R. Slonim (2013): "Is There Selection Bias in Laboratory Experiments? the Case of Social and Risk Preferences," *Experimental Economics*, 16, 372–382.

Cobo-Reyes, R. and N. Jiménez (2012): "The Dark Side of Friendship: 'envy'," *Experimental Economics*, 15, 547–570.

Coffman, K. B. (2014): "Evidence on Self-Stereotyping and the Contribution of Ideas," *Quarterly Journal of Economics*, 129, 1625–1660.

Coffman, L. C. (2011): "Intermediation Reduces Punishment (and Reward)," *American Economic Journal: Microeconomics*, 3, 77–106.

Cohn, A., J. Engelmann, E. Fehr, and M. A. Maréchal (2015): "Evidence for Countercyclical Risk Aversion: an Experiment with Financial Professionals," *American Economic Review*, 105, 860–885.

Cojoc, D. and A. Stoian (2014): "Dishonesty and Charitable Behavior," *Experimental Economics*, 17, 717–732.

Comerton-Forde, C. and T. J. Putniņš (2011): "Pricing Accuracy, Liquidity and Trader Behavior with Closing Price Manipulation," *Experimental Economics*, 14, 110–131.

Coricelli, G., M. Joffily, C. Montmarquette, and M. C. Villeval (2010): "Cheating, Emotions, and Rationality: An Experiment on Tax Evasion," *Experimental Economics*, 13, 226–247.

Cramton, P., E. Filiz-Ozbay, E. Y. Ozbay, and P. Sujarittanonta (2012): "Discrete Clock Auctions: An Experimental Study," *Experimental Economics*, 15, 309–322.

Cubitt, R. P., M. Drouvelis, and S. Gächter (2011): "Framing and Free Riding: Emotional Responses and Punishment in Social Dilemma Games," *Experimental Economics*, 14, 254–272.

Davis, D., A. Ivanov, and O. Korenok (2016): "Individual Characteristics and Behavior in Repeated Games: An Experimental Study," *Experimental Economics*, 19, 67–99.

Davis, L. R., B. P. Joyce, and M. R. Roelofs (2010): "My Money or Yours: House Money Payment Effects," *Experimental Economics*, 13, 189–205.

De Clippel, G., K. Eliaz, and B. Knight (2014): "On the Selection of Arbitrators," *American Economic Review*, 104, 3434–3458.

de Oliveira, A. C., R. T. Croson, and C. Eckel (2015): "One Bad Apple? Heterogeneity and Information in Public Good Provision," *Experimental Economics*, 18, 116–135.

Dechenaux, E., S. D. Mago, and L. Razzolini (2014): "Traffic Congestion: An Experimental Study of the Downs-Thomson Paradox," *Experimental Economics*, 17, 461–487.

Duffy, J. and J. Lafky (2016): "Birth, Death and Public Good Provision," *Experimental Economics*, 19, 317–341.

Eil, D. and J. M. Rao (2011): "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, 3, 114–138.

Erat, S. and U. Gneezy (2016): "Incentives for Creativity," *Experimental Economics*, 19, 269–280.

Exley, C. L. (2015): "Excusing Selfishness in Charitable Giving: The Role of Risk," *Review of Economic Studies*, 83, 587–628.

Fatas, E., M. A. Meléndez-Jiménez, and H. Solaz (2010): "An Experimental Analysis of Team Production in Networks," *Experimental Economics*, 13, 399–411.

Fehr, E., H. Herz, and T. Wilkening (2013): "The Lure of Authority: Motivation and Incentive Effects of Power," *American Economic Review*, 103, 1325–1359.

Fischbacher, U. and S. Gächter (2010): "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments," *American economic review*, 100, 541–556.

Frackenpohl, G., A. Hillenbrand, and S. Kube (2016): "Leadership Effectiveness and Institutional Frames," *Experimental Economics*, 19, 842–863.

Giamattei, M. and J. G. Lambsdorff (2015): "Balancing the Current Account: Experimental Evidence on Underconsumption," *Experimental Economics*, 18, 670–696.

Gill, D. and V. Prowse (2012): "A Structural Analysis of Disappointment Aversion in a Real Effort Competition," *American Economic Review*, 102, 469–503.

Goette, L., D. Huffman, and S. Meier (2012): "The Impact of Social Ties on Group Interactions: Evidence from Minimal Groups and Randomly Assigned Real Groups," *American Economic Journal: Microeconomics*, 4, 101–115.

Greiner, B., W. Güth, et al. (2012): "Social Communication and Discrimination: a Video Experiment," *Experimental Economics*, 15, 398–417.

Grimalda, G., A. Kar, and E. Proto (2016): "Procedural Fairness in Lotteries Assigning Initial Roles in a Dynamic Setting," *Experimental Economics*, 19, 819–841.

Grosskopf, B. and R. Sarin (2010): "Is Reputation Good or Bad? An Experiment," *American Economic Review*, 100, 2187–2204.

Heinemann, F., R. Nagel, and P. Ockenfels (2009): "Measuring Strategic Uncertainty in Coordination Games," *Review of Economic Studies*, 76, 181–221.

Hortala-Vallve, R., A. Llorente-Saguer, and R. Nagel (2013): "The Role of Information in Different Bargaining Protocols," *Experimental Economics*, 16, 88–113.

Hyndman, K., A. Terracol, and J. Vaksmann (2009): "Learning and Sophistication in Coordination Games," *Experimental Economics*, 12, 450–472.

Imas, A. (2016): "The Realization Effect: Risk-Taking After Realized Versus Paper Losses," *American Economic Review*, 106, 2086–2109.

Ismayilov, H. and J. Potters (2016): "Why Do Promises Affect Trustworthiness, or Do They?" *Experimental Economics*, 19, 382–393.

Jakiela, P. (2013): "Equity vs. Efficiency vs. Self-Interest: On the Use of Dictator Games to Measure Distributional Preferences," *Experimental Economics*, 16, 208–221.

Jordan, J., K. McAuliffe, and D. Rand (2016): "The Effects of Endowment Size and Strategy Method on Third Party Punishment," *Experimental Economics*, 19, 741–763.

Kessler, J. B. and A. E. Roth (2012): "Organ Allocation Policy and the Decision to Donate," *American Economic Review*, 102, 2018–2047.

Kogan, S., A. M. Kwasnica, and R. A. Weber (2011): "Coordination in the Presence of Asset Markets," *American Economic Review*, 101, 927–947.

Koppel, H. and T. Regner (2014): "Corporate Social Responsibility in the Work Place," *Experimental Economics*, 17, 347–370.

Kuziemko, I., R. W. Buell, T. Reich, and M. I. Norton (2014): "'last-Place Aversion': Evidence and Redistributive Implications," *Quarterly Journal of Economics*, 129, 105–149.

Larkin, I. and S. Leider (2012): "Incentive Schemes, Sorting, and Behavioral Biases of Employees: Experimental Evidence," *American Economic Journal: Microeconomics*, 4, 184–214.

Levin, D., J. Peck, and A. Ivanov (2016): "Separating Bayesian Updating from Non-Probabilistic Reasoning: An Experimental Investigation," *American Economic Journal: Microeconomics*, 8, 39–60.

Lien, J. W., J. Zheng, and X. Zhong (2016): "Preference Submission Timing in School Choice Matching: Testing Fairness and Efficiency in the Laboratory," *Experimental Economics*, 19, 116–150.

Maggian, V. and M. C. Villeval (2016): "Social Preferences and Lying Aversion in Children," *Experimental Economics*, 19, 663–685.

Mago, S. D. and E. Dechenaux (2009): "Price Leadership and Firm Size Asymmetry: An Experimental Analysis," *Experimental Economics*, 12, 289–317.

Mak, V. and R. Zwick (2014): "Experimenting and Learning with Localized Direct Communication," *Experimental Economics*, 17, 262–284.

Marzilli Ericson, K. M. and A. Fuster (2011): "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments," *Quarterly Journal of Economics*, 126, 1879–1907.

Mengel, F., E. Tsakas, and A. Vostroknutov (2016): "Past Experience of Uncertainty Affects Risk Aversion," *Experimental Economics*, 19, 151–176.

Miller, M., L. M. King, C. F. Camerer, and A. Rangel (2010): "Pavlovian Processes in Consumer Choice: The Physical Presence of a Good Increases Willingness-to-Pay," *American Economic Review*, 100, 1556–1571.

Nosenzo, D., S. Quercia, and M. Sefton (2015): "Cooperation in Small Groups: The Effect of Group Size," *Experimental Economics*, 18, 4–14.

Oprea, R. (2014): "Survival Versus Profit Maximization in a Dynamic Stochastic Experiment," *Econometrica*, 82, 2225–2255.

Osés-Eraso, N. and M. Viladrich-Grau (2011): "The Sustainability of the Commons: Giving and Receiving," *Experimental Economics*, 14, 458–481.

Riyanto, Y. E. and J. Zhang (2016): "Putting a Price Tag on Others' Perceptions of Us," *Experimental Economics*, 19, 480–499.

Rosenboim, M. and T. Shavit (2012): "Whose Money is it Anyway? Using Prepaid Incentives in Experimental Economics to Create a Natural Environment," *Experimental Economics*, 15, 145–157.

Ruud, P. A., D. Schunk, and J. K. Winter (2014): "Uncertainty Causes Rounding: An Experimental Study," *Experimental Economics*, 17, 391–413.

Rydval, O., A. Ortmann, S. Prokosheva, and R. Hertwig (2009): "How Certain is the Uncertainty Effect?" *Experimental Economics*, 12, 473–487.

Sautmann, A. (2013): "Contracts for Agents with Biased Beliefs: Some Theory and an Experiment," *American Economic Journal: Microeconomics*, 5, 124–156.

Shurchkov, O. (2013): "Coordination and Learning in Dynamic Global Games: Experimental Evidence," *Experimental Economics*, 16, 313–334.

Smith, A. (2013): "Estimating the Causal Effect of Beliefs on Contributions in Repeated Public Good Games," *Experimental Economics*, 16, 414–425.

Sprenger, C. (2010): "An Endowment Effect for Risk: Experimental Tests of Stochastic Reference Points," *Journal of Political Economy*, 123, 1456–1499.

Stöckl, T. (2014): "Price Efficiency and Trading Behavior in Limit Order Markets with Competing Insiders," *Experimental Economics*, 17, 314–334.

Tufano, F. (2010): "Are 'True' Preferences Revealed in Repeated Markets? An Experimental Demonstration of Context-Dependent Valuations," *Experimental Economics*, 13, 1–13.

Vieider, F. M. (2011): "Separating Real Incentives and Accountability," *Experimental Economics*, 14, 507–518.

Visser, M. S. and M. R. Roelofs (2011): "Heterogeneous Preferences for Altruism: Gender and Personality, Social Status, Giving and Taking," *Experimental Economics*, 14, 490–506.

References

Simonsohn, U., L. D. Nelson, and J. P. Simmons (2014): "*P*-Curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General*, 143, 534–547.

Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr. (1949): *The American soldier: Adjustment during army life*, vol. 1 of *Studies in social psychology in World War II,*, Princeton, NJ: Princeton University Press.