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

Taisuke Imai Klavdia Zemlianova Nikhil Kotecha
Colin F. Camerer *

January 14, 2020

Abstract

Scientific conclusions depend on how data are collected and analyzed, and whether norms and editorial practices suppress false positives. Poor replicability of results in some areas of medicine and psychology have raised concerns about how widespread such bad results might be in other areas of science. We analyzed laboratory experiments in economics published in seven leading journals between 2009 and 2016 using the p-curve method, which looks at the frequencies of reported p-values in equal-sized bins, spanning relatively strong ($p < 0.01$) and marginal ($0.04 < p < 0.05$) results. The observed p-curve is strongly right-skewed, indicating that p-hacking does not appear to be common in laboratory experimental economics.

JEL code: A11, C12
Keywords: false positives, p-hacking, experimental economics

*Imai: LMU Munich. Zemlianova: Northwestern University. Kotecha: Columbia University. Camerer: California Institute of Technology. Imai acknowledges financial support by the Deutsche Forschungsgemeinschaft through CRC TRR 190. Zemlianova acknowledges support from the Caltech Summer Undergraduate Research Fellowship (SURF) Program. Corresponding author E-mail: camerer@hss.caltech.edu.

1 Introduction

In the last decade, scientists have become more concerned about the quality of conclusions from cumulated evidence. This rising concern has also been accompanied by active steps— more actual replications, evidence of predictability of replication, and changes in scientific and editorial practice.

For example, the *American Economic Review* proceedings issue published several papers about replication in economics recently. Most papers noted the importance of replication (Höffler, 2017), and that rates of actual replication are low to medium (Berry et al., 2017; Duvendack et al., 2017; Sukhtankar, 2017). Others called for enhanced publication of replications in economics (as has already become common in psychology) and citation of those replications along with original results (Coffman et al., 2017). One author felt robustness-type reproduction of results works well in labor economics (Hamermesh, 2017). Others noted that meta-analysis is useful but requires judgment (Anderson and Kichkha, 2017).

Scientists in other fields have developed analytical methods to detect whether results that are unlikely to replicate. In this paper, we describe and apply one such method called "p-curve". The idea is simple: Suppose that scientists, or referees and editors, attach special importance to statistical results which pass a particular level of type I error (most commonly, five percent). If publications in peer-reviewed journals are important for scientists' careers, then some scientists will be motivated to "p-hack" to push p-values below the magic threshold. P-hacking methods include collecting more data only in the hope of lowering p-values, selectively discarding outliers, focusing attention on unpredicted partial results, using different statistical tests and reporting only the ones that 'worked' (the "file drawer" problem, Rosenthal, 1979).

The p-curve, introduced by Simonsohn et al. (2014), is a simple method to detect p-hacking in an aggregate body of reported evidence. The p-curve is the frequency distribution of reported p-values in equal-sized numerical intervals, such as $[0, 0.01]$, $(0.01, 0.02]$, ..., and $(0.04, 0.05]$. If there is no true effect and no p-hacking, then the expected frequencies of p-values in all those intervals should be the same (though the actual frequences will differ a bit due to sampling error). If there is p-hacking, however, then there will be disproportionately many results in the interval just below the crucial threshold, $(0.04, 0.05]$. And if there is no p-hacking and genuine effects exist and are measured with powerful tests, there will be disproportionately many p-values in lower categories.

We report p-curves for a sample of 105 published laboratory experiments in economics. The main result is that p-curves indicate genuine effects that are adequately powered. There are none

of the hallmarks of p-hacking that have been shown in other studies reviewed below.

The motivation for our analysis is occasional debate about the quality of lab data in experimental economics. We will briefly describe some arguments for, and against, the hypothesis that experimental economics might produce results that are systematically inflated, and are therefore unlikely to replicate well.

Arguments that results are inflated. Concerns about p-hacking are not new in economics. An editor of the prestigious *Journal of Political Economy*, Feige (1975) worried that:

> "[...] current journal editorial policies have undoubtedly contributed to (1) an incentive to pursue search procedures for statistically significant results which are spurious as often reported, insofar as they take no account of pretest bias; (2) an incentive for less than candid reporting of intermediate results which could highlight the lack of robustness of statistical tests to alternative model specifications and applications of alternative econometric techniques; (3) an underrepresentation of 'negative' results which could otherwise signal empirically anomalous results leading to the rejection of currently maintained hypotheses; and (4) an unnoticed proliferation of published Type 1 errors." (pp. 1292-93)

Concerns about a particular type of p-hacking in experimental economics have also been raised by Roth (1994):

> "[...] I once had the opportunity to hear one experimental economist chide another for having reported that a certain kind of market did not always yield equilibrium behavior. He felt that perhaps a premature negative result had been reported. He went on to say that, in his own research, when he found in an experiment that some economic institution 'didn't work,' he first tried rewriting the instructions to make sure that they hadn't contributed to the negative result, and if that didn't fix the problem he would try changing the mechanics of the experiment. Often, he said, that fixed the problem. *Left unstated was that this search for conditions that would yield the desired result was not reported in the papers that resulted from this activity*, which simply presented, as if they were independent experiments, trials that had 'worked.' " (p. 283, footnote 5; italics ours)

The implication in this essay is that leaving the sequence of instruction rewriting, changes in experimental "mechanics", etc., unreported in a published paper could inflate the size and robustness of the published effects. In the particular case described, the implication is that the published

evidence could be inflated, because pilot experiments (which found the opposite conclusion under different instructions and mechanics) were not reported.

It is, of course, difficult to get hard, conclusive data on the prevalence of p-hacking. The data that are available are quite similar across academic fields and are cause for concern.

John et al. (2012) conducted a survey of experimental psychologists about "questionable research practices" (QRPs) which range from fabrication, various types of p-hacking, excluding undesirable results, 'coercive' citation (to previous papers in the target journal), and accepting referee comments one disagrees with. They found low rates of admitted fabrication (less than 5%), but rates of other QRPs from 40 to 80%.

There are two large-scale surveys of QRPs by economists. List et al. (2001) surveyed 1,000 attendees at the American Economic Association meetings 1998 (response rate of 23%) about fabrication and four minor QRPs (e.g., submitting an article simultaneously to multiple journals, against journal policy, including an undeserving coauthor). The rates were about five and 10% respectively. (And as in all surveys of this form, respondents think their colleagues commit QRPs about 50% more often.) Necker (2014) surveyed 2,520 members of the European Economic Association (complete response rate of 16.9%) online about a longer list of QRPs based on John et al. (2012). She reports QRP prevalence from 20 to 59% and admitted fabrication of 2.5%.

A general approach to why replicability might be poor in economics experiments was articulated by Maniadis et al. (2014). While their critique clearly applies to all empirical work of any type, they directed it specifically at laboratory experiments. They present some algebra showing how test low power, experimenter bias, and prior probability could conceivably create false positives in experimental economics.

Their analysis is a useful start but does not include any roles for peer refereeing. For example, if referees condition their evaluation of surprising conclusions on priors, they may reject weak papers or insist that experimenters add data or alternative tests to correct biases. Thus, their model is a partial equilibrium one that might account for pre-publication results but is not likely to capture all the features of published results. On a broader level, a primary metric in the hiring of tenure-track faculty is the number of high-impact publications. On an institutional level, this structural incentive positively selects for— without direct strategizing on the part of the researcher— poor methods. These poor research methods create a persistence of false-positive results and a culture that esteems p-values over understanding.

Arguments that results are not inflated. The view opposite to the one described above is that practices in experimental economics are not conducive to p-hacking by scientists, or to spe-

cial reverence for p-values by referees and editors.

Since the quality and nature of replication is not often discussed openly in peer-reviewed publication, other sources of opinion become valuable. In 2015, some of us submitted a grant proposal to the Sloan Foundation to fund replications in experimental economics (leading to Camerer et al., 2016). One of the reviewers wrote:

> "The problems that the proposal identifies as in need of correction might be features of some areas of science but are not a substantial part of experimental economics. The proposal points to 'a surge of interest in how well scientific results replicate' (p. 3). However, the problems the proposal mentions are based on other sciences together with some assertions about problems found in empirical economics (econometric) practices. Strikingly absent are examples from experimental economics."

This reviewer is implying that poor replicability and p-hacking do not occur frequently in experimental economics.

Another argument in this optimistic direction is that experimental economists have, from the start, been more cautious about transparency and robustness, because they faced a more skeptical audience, than in other social sciences where the experimental method was immediately accepted. Authors were eager to be transparent to bolster credibility of the newly-emerging practice of experimentation in economics, and to make replication easier.

A third argument against likely inflation of published results is that many economics experiments do not test or discover complicated directional hypotheses, which are surprising. Failures of speculative directional tests, or searches for interactions, might be what dooms papers to a dark fate in the proverbial file drawer.

Instead, many economics experiments predict point estimates of behavioral observables from design variables and theory, or simple directional main effects. Examples included price and volume predictions in double auction designs, the slopes of bid functions in auctions, price paths in artificial stock markets, and the influence of time horizon and payoffs on cooperation in repeated prisoners' dilemma games.

The last argument why experimental economics results are unlikely to be inflated is that interactive classroom experiments are often used to teach introductory, principles classes. If experiments failed to replicate, these classroom demonstrations would be duds; the instructor would expect a result that would not always happen, undermining her planned lesson. In general, experimental demonstrations are not duds. Other studies indicate that the ability of experimental economics results to replicate appears to be solid.

Related research. In ongoing research on scientific reproducibility, a natural concern is that an obsession with statistically significant results has led to selective reporting practices and substantial bias (perhaps even in editorial practices, not just in what scientists discover and choose to report). Several studies have examined clustering around key significance thresholds (e.g., Gerber and Malhotra, 2008) or the entire distribution of test statistics (Brodeur et al., 2016). Usually there is a dip in below-threshold reporting.

The p-curve method identifies distortions in the distribution of p-values below the 0.05 threshold to search for "p-hacking" (too many p-values just under the 0.05 threshold). The method has been used in the medical literature (Jager and Leek, 2014) and in a range of disciplines in science (Head et al., 2015) and social science (Tanner, 2015). For instance, Simmons and Simonsohn (2017) use the p-curve method to explore the 33 papers reviewed in the power-posing work of Carney et al. (2015). Simmons and Simonsohn (2017) find the distribution of p-values from those 33 papers to be indistinguishable from a distribution with an average effect size of zero, and selective reporting as the source of the published significant results. Rand (2016) use the p-curve (among other tests) to examine 19 studies which investigated the role of intuition and deliberation applying cognitive-processing manipulations to economic games. The p-curve indicates the presence of evidential value (i.e., people cooperated more when the use of intuition was promoted over deliberation).

2 The *P*-Curve Method

Simonsohn et al. (2014) introduced a method called *p-curve* to diagnose selective reporting in a set of statistically significant findings. According to Simonsohn et al. (2014), a set of significant findings contains *evidential value* when we can rule out selective reporting as the sole explanation of these findings. One way to test the existence of evidential value is to look at the distribution of reported p-values in a *set* of studies. A p-curve simply refers to the distribution of p-values from a set of independent findings. The fundamental idea behind the method is to make inferences from the shape of the p-curve.

The examples we present focus on low p-values (below 0.05). To improve visualization, we will usually present relative frequencies of p-values in discretized bins. A simple, important comparison is between the percentage of p-values in the interval $[0, 0.01]$, and the percentage in the interval $(0.04, 0.05]$.

The idea behind the p-curve method is simple and intuitive. In hypothesis testing, a p-value expresses the probability of obtaining data at least as extreme as the one observed (type I error)

if there is no genuine effect. If the null hypothesis is true— in other words, there is no genuine effect– then p-values based on a continuous test statistic will be uniformly distributed regardless of the sample size of the observations. This implies that, under the null hypothesis of no-effect, the p-curve will be flat.

Now suppose the alternative hypothesis is true— in other words, a nonzero hypothesized effect does exist. Then the distribution of p-values from independent tests will *not* be uniform. The precise distribution depends on the sample size and the power of a study, but in all common cases (e.g., a parametric t-test assuming normally-distributed data), the p-curve becomes right-skewed: That is, the percentage of p-values in the very low interval $[0, 0.01]$ will be higher than the percentage in the just-below 0.05 interval $(0.04, 0.05]$. Intuitively, if there is an effect and a set of designs are well-powered to detect it, there should be a lot more very strong, low p-values than marginal results just below the 0.05 norm.

If only significant hypothesized results can be published, publication-minded researchers could engage in two types of questionable research practices that undermine reproducibility. One is to put insignificant results in the proverbial file-drawer. Another is to choose sample sizes endogenously, discard outliers, shift attention to whether hypothesized effects are present in subsamples chosen post-hoc, or selectively include covariates. Whether implicit or explicit, the purpose of these practices is to produce a significant result. Simonsohn et al. (2014) called such behavior *p-hacking*. To be clear, p-hacking is an overproduction of marginally significant results. Even in the absence of p-hacking, file-drawer burial and editorial rejection of insignificant results can also result in an extraordinary number of p-values just below the threshold, which means that the shape of the p-curve would become left-skewed. This is because researchers are likely to stop further investigation once they obtain significance, resulting in a disproportionately large proportion of high (i.e., close to 0.05) p-values. In reality the story is a little more complicated, since the shape of a p-curve would depend on the power of the study and the intensity of p-hacking. Studies with low power and intense enough p-hacking could produce left-skewed p-curves. In addition, studies with genuine effects and sufficient power, and mild p-hacking, could also produce right-skewed p-curve.

Based on those observations, Simonsohn et al. (2014) proposed to test for skewness of the observed p-curve to examine whether a set of studies contain evidential value. See online supplementary material A for details.

3 Data

We looked at experimental studies published in the following seven journals between 2009 and 2016: The *American Economic Journal: Microeconomics* (AEJ:Mic), the *American Economic Review* (AER), *Econometrica* (ECMA), *Experimental Economics* (EE), the *Journal of Political Economy* (JPE), the *Quarterly Journal of Economics* (QJE), and the *Review of Economic Studies* (REStud). We examined only laboratory experiments and omitted natural experiments, field and online experiments, and tests on existing datasets or meta-analyses. We hereafter call six journals other than EE collectively "Top 5+1".

The first major challenge is identification of the main hypothesis of interests. Selecting from the many statistics reported is not a simple task (unless they are clearly stated as the main tests). The selection is inherently subjective. To maintain consistency, we established a procedure listed below for inclusion and for further categorization of reported test statistics.

From each paper, the main hypothesis was determined to be what the author(s) claimed to be the main result/finding in the abstract, results, discussion and/or conclusion sections, usually in the form of "the main result is..." or "contribution is...".[1] If the author(s) did not explicitly state what the main hypothesis was, then judgment was based on what result the abstract, results, discussion and/or conclusion sections focused on. In cases in which papers were unclear or implicit about their main hypothesis, we labeled "potentially main" hypothesis with two indices: *study number* and *hypothesis number*. The former refers to the specific study in the paper and the latter captured if the authors conducted several tests to check the hypothesis.[2]

Another challenge in assembling a proper dataset for p-curve analysis is the correlation among test statistics. For example, if the main hypothesis was tested using t-tests on two different but correlated measures, there is no a-priori strong reason to include one test and exclude the other. We included all relevant tests in the dataset but picked one for the main analysis. In later robustness checks, we constructed "bootstrap" datasets by picking one test randomly from each paper.

[1] P-values from robustness checks section, model verification, confirming past findings were not counted as main result and thus omitted from our data collection.

[2] Some studies tested the main hypotheses using the fact that the relevant tests were *not* significant (e.g., to show that there is no gender effect). These data do not help identify the shape of the p-curve so they are excluded.

4 Results

The dataset consists of the total of 237 significant ($p < 0.05$) test statistics taken from 105 papers (online supplementary material C). There are 10 papers from AEJ:Mic, 20 from AER, three from ECMA, 58 from EE, two from JPE, and seven from QJE, and five from REStud.[3]

The analysis was implemented using p-curve app 4.05 (March 2017 version).[4] Figure 1A presents the result—The solid line represents the observed p-curve, the dashed line represents the p-curve one would expect if studies included in the analysis were powered at 33%, and the dash-dot line represents the p-curve one would expect under no evidential value. In order to maintain the validity of p-curve analysis, we followed the suggestion made by Simonsohn et al. (2014) and included only one p-value from each study. If there are multiple "potential main tests," we picked one randomly. As we described in previous sections, a left-skewed p-curve is expected if the researchers in the field of experimental economics were chasing small (or zero) effects and actively p-hacked. The curve and associated statistical analyses, however, indicated the presence of strong evidential value (absence of strong p-hacking in the literature as a whole) because left-skewness of the p-curve is rejected (Binomial test $p < 0.0001$; continuous test with Stouffer method yields $p < 0.0001$ for both full and half p-curves).

Similarly, Figure 1BC compare p-curves from EE and Top 5+1 publications. The two curves have similar shape, and left-skewness of the p-curve is strongly rejected in each of these plots.

Finally, Figures B.3-B.5 in online supplementary material present p-curves splitting samples by median citation counts. In each figure, the left panel (A) corresponds to the below-median group and the right panel (B) corresponds to the above-median group. In all of these figures left-skewness of the p-curve is rejected.

As we discussed in Section 3, we made a judgment about which one test statistic to include in the analysis to maintain independence of observations in the p-curve analysis. Our previous results might be driven by the selection of test statistics we made when the "main" ones were not clear. In order to asses the robustness of our p-curve analysis presented above, we employed a bootstrap approach to construct confidence bands around p-curves.

In each bootstrap iteration, we randomly picked one significant p-values from the set of

[3]Table B.1, Figures B.1, and B.2 in online supplementary material show the distributions of the number of papers published in each year and the distributions of citation counts collected from Google Scholar between April and June 2017.

[4]This version of the p-curve method accepts only parametric tests using z, t, F, and χ^2 statistics and Pearson's correlation coefficient r. We therefore omitted papers which tested their main hypotheses using only nonparametric tests such as Wilcoxon signed-rank test, Mann-Whitney U, Kolmogorov-Smirnov test, and others.

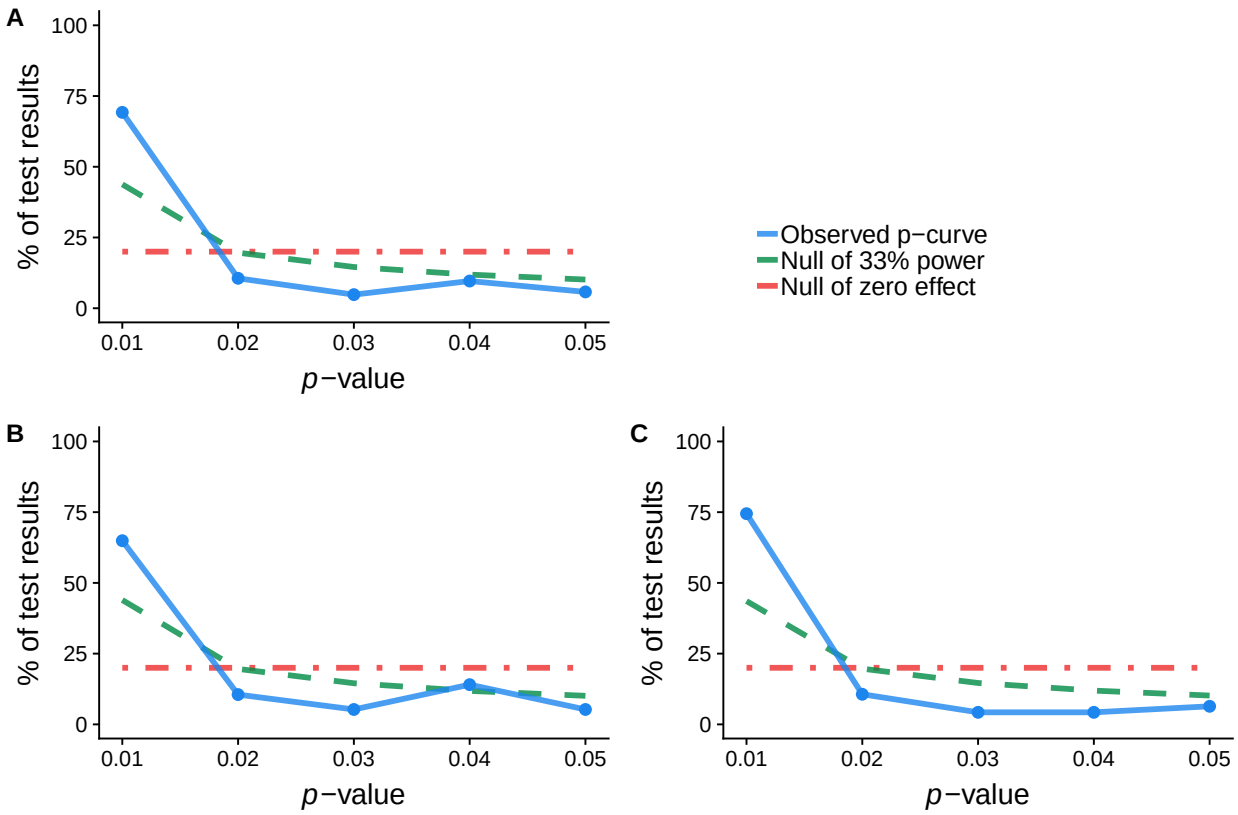

FIGURE 1: *P*-curve from: (A) all studies in the sample, (B) EE, and (C) Top 5+1.

recorded p-values associated with "potentially main" tests. After iterating 50 times, we calculated the mean p-curve and ±1 standard deviation around it. Figure 2 presents the result. Panel A confirms right-skewness of the p-curve we documented above. Similarly, panel B confirms that distributions of significant p-values reported in EE and Top 5+1 are not that different. Taken together, our main results are not driven by our subjective judgment of the main hypothesis in each study.

5 Conclusion

This paper is about inferring general quality of data in laboratory experimental economics from published frequencies of p-values. Speculations that experimental practices generate low reproducibility of results have been aired occasionally many years ago, and also more recently.

We bring systematic analysis to bear on this question by examining all laboratory experiments in economics published in seven leading journals from 2009 to 2016. The p-curve for these

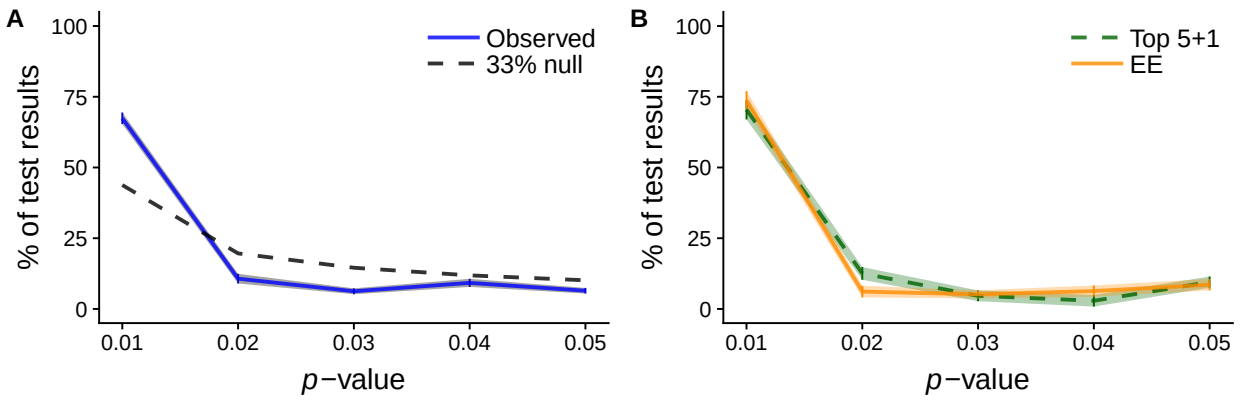

FIGURE 2: Bootstrapped p-curve: (A) all journals and (B) EE vs. Top 5+1. Error bands represent ±1 standard deviation around the mean p-curve.

published results are strongly right-skewed; that is, there are many more results with very low p-values (below 0.01) than results just below a $p < 0.05$ threshold. This pattern is consistent with substantial genuine effects, measured by high-power studies, without too much p-hacking. To be crystal clear, it is impossible to tell whether there is no p-hacking at all, because the p-curve will be right-skewed, even if there is p-hacking, if effects and power are large enough. In any case, there are no aggregate signs of routine practices creating false positives in published papers in laboratory experimental economics.

References

ANDERSON, R. G. AND A. KICHKHA (2017): "Replication, Meta-Analysis, and Research Synthesis in Economics," *American Economic Review: Papers and Proceedings*, 107, 56–59.

BERRY, J., L. C. COFFMAN, D. HANLEY, R. GIHLEB, AND A. J. WILSON (2017): "Assessing the Rate of Replication in Economics," *American Economic Review: Papers and Proceedings*, 107, 27–31.

BRODEUR, A., M. LÉ, M. SANGNIER, AND Y. ZYLBREBERG (2016): "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, 8, 1–32.

CAMERER, C. F., A. DREBER, E. FORSELL, T.-H. HO, J. HUBER, M. JOHANNESSON, M. KIRCHLER, J. ALMENBERG, A. ALTMEJD, T. CHAN, E. HEIKENSTEN, F. HOLZMEISTER, T. IMAI, S. ISAKSSON, G. NAVE, T. PFEIFFER, M. RAZEN, AND H. WU (2016): "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351, 1433–1436.

Carney, D. R., A. J. C. Cuddy, and A. J. Yap (2015): "Review and Summary of Research on the Embodied Effects of Expansive (vs. Contractive) Nonverbal Displays," *Psychological Science*, 26, 657–663.

Coffman, L. C., M. Niederle, A. J. Wilson, et al. (2017): "A Proposal to Organize and Promote Replications," *American Economic Review: Papers and Proceedings*, 107, 41–45.

Duvendack, M., R. Palmer-Jones, and W. R. Reed (2017): "What Is Meant by "Replication" and Why Does It Encounter Resistance in Economics?" *American Economic Review: Papers and Proceedings*, 107, 46–51.

Feige, E. L. (1975): "The Consequences of Journal Editorial Policies and a Suggestion for Revision," *Journal of Political Economy*, 83, 1291–1296.

Gerber, A. and N. Malhotra (2008): "Do Statistical Reporting Standards Affect What is Published? Publication Bias in Two Leading Political Science Journals," *Quarterly Journal of Political Science*, 3, 313–326.

Hamermesh, D. S. (2017): "Replication in Labor Economics: Evidence from Data, and What It Suggests," *American Economic Review: Papers and Proceedings*, 107, 37–40.

Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions (2015): "The Extent and Consequences of P-Hacking in Science," *PLoS Biology*, 13, e1002106.

Höffler, J. H. (2017): "Replication and Economics Journal Policies," *American Economic Review: Papers and Proceedings*, 107, 52–55.

Jager, L. R. and J. T. Leek (2014): "An Estimate of the Science-Wise False Discovery Rate and Application to the Top Medical Literature," *Biostatistics*, 15, 1–12.

John, L. K., G. Loewenstein, and D. Prelec (2012): "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling," *Psychological Science*, 23, 524–532.

List, J. A., C. D. Bailey, P. J. Euzent, and T. L. Martin (2001): "Academic Economists Behaving Badly? A Survey on Three Areas of Unethical Behavior," *Economic Inquiry*, 39, 162–170.

Maniadis, Z., F. Tufano, and J. A. List (2014): "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects," *American Economic Review*, 104, 277–290.

Necker, S. (2014): "Scientific Misbehavior in Economics," *Research Policy*, 43, 1747–1759.

Rand, D. G. (2016): "Cooperation, Fast and Slow: Meta-Analytic Evidence for a Theory of Social Heuristics and Self-Interested Deliberation," *Psychological Science*, 27, 1192–1206.

Rosenthal, R. (1979): "The File Drawer Problem and Tolerance for Null Results," *Psychological Bulletin*, 86, 638–641.

Roth, A. E. (1994): "Lets Keep the Con out of Experimental Econ.: A Methodological Note," *Empirical Economics*, 19, 279–89.

Simmons, J. P. and U. Simonsohn (2017): "Power Posing: *P*-Curving the Evidence," *Psychological Science*, 28, 687–693.

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.

Sukhtankar, S. (2017): "Replications in Development Economics," *American Economic Review: Papers and Proceedings*, 107, 32–36.

Tanner, S. (2015): "Evidence of False Positives in Research Clearinghouses and Influential Journals: An Application of P-Curve to Policy Research," *Observational Studies*, 1, 18–29.