Brodeur, Abel; Cook, Nikolai M.; Hartley, Jonathan S.; Heyes, Anthony

Working Paper

Do Pre-Registration and Pre-analysis Plans Reduce p-Hacking and Publication Bias?

GLO Discussion Paper, No. 1147

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Brodeur, Abel; Cook, Nikolai M.; Hartley, Jonathan S.; Heyes, Anthony (2022) : Do Pre-Registration and Pre-analysis Plans Reduce p-Hacking and Publication Bias?, GLO Discussion Paper, No. 1147, Global Labor Organization (GLO), Essen

This Version is available at:
https://hdl.handle.net/10419/262738

Do Pre-Registration and Pre-analysis Plans Reduce p-Hacking and Publication Bias?*

Abel Brodeur Nikolai M. Cook Jonathan S. Hartley
Anthony Heyes

August 3, 2022

Abstract

Randomized controlled trials (RCTs) are increasingly prominent in economics, with pre-registration and pre-analysis plans (PAPs) promoted as important in ensuring the credibility of findings. We investigate whether these tools reduce the extent of p-hacking and publication bias by collecting and studying the universe of test statistics, 15,992 in total, from RCTs published in 15 leading economics journals from 2018 through 2021. In our primary analysis, we find no meaningful difference in the distribution of test statistics from pre-registered studies, compared to their non-pre-registered counterparts. However, pre-registered studies that have a complete PAP are significantly less p-hacked. These results point to the importance of PAPs, rather than pre-registration in itself, in ensuring credibility.

KEYWORDS: Pre-analysis plan - Pre-registration - p-Hacking - Publication bias - Research credibility

JEL CODES: B41, C13, C40, C93.

*Authors: Brodeur: University of Ottawa and IZA. E-mail: abrodeur@uottawa.ca. Cook: Wilfrid Laurier University. ncook@wlu.ca. Hartley: Stanford University. hartleyj@stanford.edu. Heyes: University of Birmingham. aheyes@uottawa.ca. We thank Isaiah Andrews, Nina Buchmann, Arun Chandrasekhar, Pascaline Dupas, Marcel Fafchamps, John List, Steve Levitt, Eva Lestant, Neil Malhotra, Melanie Morten and Muriel Niederle for helpful comments. We thank Abigail Marsh and Susan Price for research assistance. This pre-analysis plan was written at the end of data collection of test statistics but prior to (i) coding pre-registration status and (ii) coding whether the presence of a complete PAP. We had not cleaned the data nor conducted any empirical analysis at the time of pre-registration. Heyes acknowledges financial support from the Canada Research Chair programme. Errors are ours.

1 Introduction

Randomized controlled trials (RCTs) have become increasingly popular in economics and the social sciences more broadly, in recent years. By combining deliberate randomization of treatment with observation of subjects in naturally-occurring settings, the RCT is often regarded as the ideal or 'gold standard' methodology for causal inference (Ravallion (2020)). This increased prominence since the start of the millennium has coincided with a broader push to normalize practices of open science and research transparency aimed at bolstering the credibility of empirical research in the discipline. In the particular case of RCTs, this push has promoted the importance of pre-analysis plans (PAP) and pre-registration of research projects. Reflecting this, since the American Economic Association (AEA) launched the AEA RCT Registry in 2013, Banerjee et al. (2020) report that over 2,165 trials have been registered (as of January 2020) and, of those, 1,153 were pre-registered.

In this paper, we investigate the relationship between pre-registration, pre-analysis plans and statistical significance. This research question is of particular interest since proponents of pre-registered PAPs point to their scope to reduce p-hacking (i.e., manipulation and/or selective reporting of results' p-values) and publication bias (i.e., the statistical significance of results in a study influencing the decision to publish), making the published body of research more credible or trust-worthy.[1] Pre-registration and PAPs are distinct and separable things. Basic pre-registration frequently requires only rather general posting that a study is to take place, whereas a pre-analysis plan requires authors to register much more explicitly which hypotheses are to be tested and how. It is thus hoped that pre-registration of a PAP will provide the researcher with less 'wiggle' room after data has been collected, so reduce the extent of p-hacking by limiting the scope for a researcher to change their analytical or reporting choices once statistical significance has been observed. Pre-registered PAPs may also reduce publication bias if

[1] See Banerjee et al. (2020), Coffman and Niederle (2015) and Olken (2015) for thoughtful discussions of the advantages and disadvantages of PAPs.

it implies that null results are more likely to be reported.

We provide what we believe to be the first systematic investigation of whether PAPs and pre-registration reduce p-hacking and publication bias in RCT-based research published in a set of high-profile economics journals. To do this we harvest and analyze the universe of hypothesis tests drawn from randomized control trials (RCTs) published in 15 leading journals in the years 2018 through 2021. The analysis includes 314 journal articles and 16,000 test statistics.

As a first step, we show that the use of PAPs and pre-registration in economics increased substantially over our study period. Defining a pre-registered RCT as a study that was registered before its trial end date listed in a registry, we find that less than 15% of RCTs in our sample were pre-registered in 2018 compared to 40% in 2021 (Figure 1).

We next investigate the relationship between articles and authors' characteristics and the use of pre-registration. This may be important if, for example, "elite" researchers are more/less likely to pre-register their RCTs (Christensen et al. (2020)) and more/less likely to p-hack. We find little evidence that more experienced and "elite" scholars are particularly prone to adopt and use pre-registration. We do, however, provide evidence that pre-registration rates vary widely across journals - with the very top journals having particularly high rates of pre-registration.

For our p-hacking and publication bias analyses, we first plot and compare visually the distribution of test statistics for pre-registered and non-pre-registered RCTs. We then test more formally using a series of methodologies. First, we apply caliper tests as in Gerber and Malhotra (2008). Caliper tests focus on the local distribution of z-statistics within a narrow band either side of conventional significance thresholds. Second, we extend and apply the methodology in Brodeur et al. (2016) and Brodeur et al. (2020) to quantify the excess (or dearth) of z-statistics over significance regions by comparing the observed distribution of test statistics for pre-registered and non-pre-registered RCTs against a counterfactual distribution that we would expect to observe absent of p-hacking and publication bias. Third, we

apply the battery of p-hacking tests developed and proposed in Elliott et al. (2022); those based on the non-increasingness of the p-curve and those testing for discontinuities. Fourth, we apply the method recently proposed by Andrews and Kasy (2019) to measure, and compare, the extent of publication bias for pre-registered and non-pre-registered RCTs.

Across the analyses we find little or no evidence to suggest that pre-registration *in itself* reduces p-hacking or publication bias.

This result may surprise empirical researchers many of whom, anecdotally at least, seem to attach significant weight to the pre-registration status of a study. A number of rationales could be proposed for the lack of effect including: (1) that RCTs, irrespective of pre-registration status, have been shown to be less p-hacked than other empirical methods commonly applied in economics; and (2) that the requirements of pre-registration, as practised by economists, are not sufficiently comprehensive to provide adequately tight constraints on research practices.

To try to tease out the mechanism(s) at play we report the results of two further exercises. First, we show that authors of pre-registered RCTs report significantly more test statistics than do their non-pre-registered RCTs, consistent with researchers following (at least to some extent) a commitment to report their pre-specified specifications. Second, we explore the additional role of pre-registration including a PAP. We find that inclusion of a PAP in the pre-registration is associated with a decrease in reduction in p-hacking and publication bias. Relatedly, pre-registered RCTs that provide an explicit discussion of statistical power/sample size appear less prone to p-hacking than those that do not provide such a discussion.

We conclude that the extent of p-hacking published RCT-based research in economics is relatively small - other non-experimental methods that have been shown to be more compromised might benefit more from the use of pre-registration and PAP (Burlig (2018)). Nonetheless, we document that there is room for improvements even for RCTs, with pre-registration only appearing to improve the credibility of results when that pre-registration includes a detailed PAP. Given that the processes

of pre-registration and pre-analysis are costly whether the benefits in improved credibility justify the costs are not questions that we address here.

Our study contributes to the growing and important literature that develops and tests the effectiveness of new open science practices (Brandon and List (2015); Blanco-Perez and Brodeur (2020); Brodeur et al. (2022); Butera et al. (2020); Camerer et al. (2019); Casey et al. (2012); Christensen et al. (2019); Drazen et al. (2021)). Three studies are especially pertinent. Franco et al. (2016) use the Time-sharing Experiments for the Social Sciences (TESS) database as a way to explore published outputs from pre-registered projects, finding that about 40% of studies fail to report all experimental conditions and about 70% of studies do not report all of the outcome variables included in the questionnaire. Ofosu and Posner (2021) analyze the content of 195 PAPs registered on the Evidence in Governance and Politics (EGAP) and AEA registration platforms from 2011 to 2016, and argue that PAPs are not sufficiently comprehensive to achieve their intended objectives. Last, in an unpublished study, Fang and Humphreys (2015) examine changes in the distribution of published p-values before and after the introduction of registration requirements for medical journals, and find no evidence that registration impacted the distribution of p-values near significance cut-offs. To our knowledge, we are the first to document the relationship between p-hacking, publication bias and pre-registration in the social sciences.

Our study also contributes to the much broader body of literature documenting the extent of p-hacking and publication bias in economics and related disciplines more broadly, not just with respect to RCTs (Andrews and Kasy (2019); Doucouliagos and Stanley (2013); Furukawa (2019); Gerber and Malhotra (2008); Havránek (2015); Havránek and Sokolova (2020)).[2] The most relevant studies are Brodeur et al. (2016), Brodeur et al. (2020) and Vivalt (2019) who show that p-hacking is less prevalent for papers using RCTs than those using other methods of causal

[2]See Christensen and Miguel (2018), Miguel (2021) and and Stanley and Doucouliagos (2014) for literature reviews.

inference.[3]

Section 2 provides a conceptual framework. Section 3 details the data collection. Section 4 investigates whether the likelihood of pre-registering a PAP is related to articles and authors' characteristics. In section 5, we present the distribution of test statistics by pre-registration status and formally test for the presence of p-hacking and publication bias. In Section , we explore whether pre-registration impacts reporting behavior and the role of pre-analysis plans. Last, Section 7 concludes.

2 Conceptual Framework

We first provide a brief conceptual framework and rationale for why PAPs and pre-registration might be expected to decrease p-hacking and publication bias. A more complete discussion of the advantages and disadvantages of PAP and pre-registration in economics is provided in each of Banerjee et al. (2020), Coffman and Niederle (2015), and Olken (2015).

The main advantage often claimed of pre-registering a PAP is that it reduces the prevalence of p-hacking by tying researchers' hands to the contents of their PAP. This requires the authors' pre-specification be sufficiently precise and exhaustive in setting out, in advance of examining data, the way the analysis will be conducted. A complete PAP would include the econometric specifications, outcome variables, cleaning procedures and other methodological details that could ultimately influence statistical significance. If the regressions are pre-specified and researchers report (or are required to report) all the results pre-specified, p-hacking becomes much less of a problem.[4]

A related benefit is a potential reduction in the extent of publication bias, at least at the working paper stage. By pre-registering a PAP, researchers are plausibly more likely to report null results, following their commitment in advance to so doing.

[3]Another relevant study is Scheel et al. (2021) who provide evidence that registered reports (RRs) in psychology increase reporting of "negative" (not supportive of tested hypothesis) results.

[4]Pre-registering a PAP may also be useful with respect to the relationship between a researcher and 'invested' partners, for example by insulating him from pressures to show that a program is effective (Banerjee et al. (2020)).

Pre-registration may also prevent or make harder hypothesizing after results are known, a practice often referred to as "HARKing". Pre-registration of PAPs may address this problem if researchers pre-specify the whole set of hypotheses that will be tested, as good practice would say that they should.[5]

However, pre-analysis plans and pre-registration imply costs. Writing a PAP in economics is not straightforward as there are usually multiple hypotheses, with many outcome variables of interest (Olken (2015)). There are typically many modeling and data handling choices that can feed into even an apparently simple piece of empirical research. Writing a PAP may thus be time consuming (Ofosu and Posner (2021)), particularly if it is to be sufficiently exhaustive to reduce significantly the 'wiggle room' available to researchers once they have started seeing results, and so deliver the benefits just described. This cost is plausibly particularly important for scholars with fewer resources - research time, research assistance, and so on - such as early career researchers, or those working at less research-intensive institutions (Banerjee et al. (2020)).[6]

In some cases it might also not be possible strictly to follow a PAP, especially for RCTs that are implemented in an unstable environment or over an extended period. Unanticipated issues might arise during implementation (such as high attrition or low take-up) that may also require changes to the intended analysis or research design. In addition, post-registration thinking and modeling, for example contextual insight gleaned only during the execution of a trial, may meaningfully improve the value of a paper. New data may also become available, or outcomes of interest might arise.

[5] A related advantage in some contexts is that a PAP allows researchers to increase their statistical power by using one-sided hypothesis tests, having stated a hypothesis with a 'direction'. In practice, use of one-sided tests in economics is rare. We encountered only one study mentioning this advantage and those authors nonetheless went on to report two-sided hypothesis tests, following convention.

[6] See Banerjee et al. (2020) for a discussion of the relative cost of undertaking an RCT in comparison to non-experimental work where PAPs are not advocated nor required.

3 Data

We focus on leading economic journals for the years 2018 through 2021. We select the highest 15 journals as ranked using RePEc's Simple Impact Factor (2018 Simple Impact Factor, calculated over the last ten years) excluding any journal that did not publish at least one paper using RCTs. See Table A1 for the list of journals. Our final sample includes 314 journal articles.

We began by searching the entire body of published articles for keywords related to RCTs such as 'randomization' and 'randomized'.[7] From the included articles, we collected estimates only from results tables. Following Brodeur et al. (2020) we collect only coefficients of interest, excluding constant terms, balance and robustness checks, regression controls, and placebo tests. Our final sample includes 15,992 test statistics (about 51 test statistics per article). Noting that authors and journals vary with respect to the precision with which estimates are reported, we collect all decimal places.

Articles were independently coded by two of the authors, allowing us to reproduce the work of one another and make sure we only selected coefficients of interest. Note that we collected the same test statistics for the vast majority of the articles and revisited test statistics in the small number of cases in which there was initial disagreement. We will provide a robustness check excluding the comparatively small number of test statistics for which there was disagreement.

For each test statistic, we record how it is reported (e.g., t-statistic versus coefficient and standard error). We treat coefficient and standard error ratios as if they follow an asymptotically standard normal distribution. When articles report t-statistics or p-values, we transform them into equivalent z-statistics.

Finally, we collect various contextual data. For each article, we record: the journal and year of publication; the number of authors; gender of authors; the affiliations of authors at time of publication; when and from what institution they

[7] We did not search for observational studies using a PAP as those are relatively rare (Burlig (2018)).

graduated; and whether they are editors of an academic journal at the time of publication. The latter information will be collected from author websites and curriculum vitae found online. We code top institutions using the highest rated 20 in RePec's ranking of top institutions.[8]

Registration and pre-registration are coded and defined as follows. First, we flag articles where the text of the article contains one or more of the following keywords (ignoring case and using wildcard *): `aearct*, osf*, pre-regist*, pre-regist*, pre-analysis plan, socialscienceregistry`, and `PAP`. Second, each of these articles is opened and the associated keyword's context manually read to ensure correct encoding.

We define a pre-registered RCT as a study that was registered before its trial end date listed in a registry (e.g., in the AEA RCT registry). Studies that were registered after the trial end date are counted as non-pre-registered. Figure 1 illustrates pre-registration rates over time.

As a robustness exercise we consider an alternative definition of pre-registration: RCTs for which the initial registration date is before its trial start date. This second definition codes as not pre-registered those RCTs that are pre-registered after the trial start date.

Importantly, we were blind to an article's pre-registration status when manually selecting and coding test statistics, but in most cases knew whether the study was registered on the AEA RCT registry or OSF as this is reported by the authors.

Finally, we take care to deal with some complications noted in Brodeur et al. (2016). These include re-weighting articles with relatively more/less test statistics per article, and adjusting for the rounding by authors of statistics.[9]

We present summary statistics in Tables 1 and 2. In Table 1, we report the mean

[8]The following 20 institutions will be coded as top: Barcelona GSE, Boston University, Brown, Chicago, Columbia, Dartmouth, Harvard, LSE, MIT, Northwestern, NYU, Princeton, PSE, TSE, UC Berkeley, UCL, UCSD, UPenn, Stanford, and Yale. See https://ideas.repec.org/top/top.econdept.html.

[9]The correct approach to weighting of articles with relatively more/less test statistics per article is ambiguous since the number of test statistics could be a direct outcome of pre-registering a RCT. We do not re-weight articles in our baseline analysis, but show that re-weighting as no impact on our conclusions in a set of robustness exercises.

and standard deviations for key variables. We split by pre-registration status to investigate any differences along this dimension in Table 2. The unit of observation is test statistic in both tables. In our sample, researchers have about 12 years of experience (i.e., years since PhD completion). Our categorization of institutions into top and non-top show that 44% of authors graduated from a top institution, while 25% are affiliated to a top institution. RCTs have over three authors on average and only 7% of tests statistics are in solo authored articles. About 65% of authors were editors of an academic journal at the time of publication. About 30% of articles are published in 'Top 5' journals.[10] Appendix Table A1 provides a breakdown by journal.

Around 30% of the test statistics in our sample are taken from pre-registered studies. However, pre-registration rates vary considerably across journals, from over 60% for *American Economic Review* and *Journal of Political Economy*, to less than 5% for *Econometrica*, *Journal of Finance* and *Review of Economic Studies* (Appendix Table A1). Pre-registration is also remarkably higher for articles with more authors and a higher share of authors who graduated from and are affiliated to a top institution. We formally test these differences in the next section.

Of particular interest are the American Economic Association journals, which represent roughly 30% of RCTs in our sample. As of January 2018, the American Economic Association journals required that all field experiment submissions be *registered* and assigned an AEARCT number. Nonetheless, we find that only 35% of test statistics in the AEA journals are in articles that were pre-registered.

4 Determinants of Pre-Registration

We first investigate whether the propensity for a study to be pre-registered (or having a PAP) is related to article and author characteristics. Christensen et al. (2020) hypothesize that "elite" scholars may be particularly influential and supportive in adopting open science practices. They also plausibly have easier access to resources

[10]The Top 5 journals in economics are the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*.

(research assistants, etc..) that reduce the opportunity cost of such practices. In a related study, Ofosu and Posner (2020) compare the publication rates of experimental NBER working papers published between 2011 and 2018 with and without PAPs. They find that articles with PAPs are slightly less likely to be published, but other things equal more likely to land in Top 5 journals.

We rely on probit regressions that include our contextual data simultaneously. The equation is:

$$P(Preregist_{iaj}) = \Phi(\alpha + \beta_j + \gamma X_{ia}), \tag{1}$$

where $Preregist_{iaj}$ is a dummy variable for whether test i in journal article a in journal j is pre-registered (or having a PAP in a secondary analysis). X_{ia} includes a dummy variable for whether the submission is solo authored and the following author-level characteristics aggregated to the paper-level: average years since PhD, average years since PhD squared, average PhD institutional rank, average institutional rank, share of female authors, and an indicator for whether at least one of the authors was an editor of an economics journal at the time of publication. We include year fixed effects and, in some models, we add dummy variables for Top 5 journals and the three American Economic Association (AEA) journals. As noted, we rely on probit models throughout and report marginal effects. We cluster the standard errors at the journal article-level.

We report the results in Table 3. In column 1, we include all our variables simultaneously with year fixed effects. In column 2, we add dummies for the AEA and Top 5 journals. In column 3, we add journal fixed effects. In columns 4 to 6, we replicate columns 1 to 3 but use the inverse of the number of tests presented in the same article to weight observations.

We find that test statistics in articles published in 2021 are statistically significantly more likely to be pre-registered than in 2018. The point estimates are very large and significant in all columns. We also find that solo authored articles and articles with a higher share of women are significantly less likely to be pre-registered.

Experience does not seem to play an important role in determining pre-registration

nor does current affiliation ranking. We find some weak evidence that authors graduating from top institutions are more likely to pre-register, as the estimates are statistically significant at the 10% level solely in one column, though would not want to over-interpret that. Overall, we do not find much evidence supporting the idea that "elite" scholars are particularly prone to adopt and use pre-registration. Rather, we find that journal ranking plays a more important role with RCTs in Top 5 journals being 25 percentage points more likely to be pre-registered. We also document large differences, perhaps surprisingly large, in pre-registration rates across journals (see Appendix Table A1).

To sum up, our results suggest that articles' characteristics play a large role in explaining pre-registration. We are therefore careful when interpreting the relationship between pre-registration and p-hacking as articles' characteristics could also relate to p-hacking behavior. In what follows, we rely on several methods to measure p-hacking, including caliper tests which allow us to control for confounders.

5 Pre-Registration, p-Hacking and Publication Bias

In what follows, we describe our graphical and formal analyses to detect and measure p-hacking and publication bias. We rely on several methods comparing the distribution of test statistics by pre-registration status. None of the methodologies delivers a definitive proof of the impact of pre-registration on p-hacking and publication bias, but taken as a whole, we do believe that most readers will find the congruence in results across the different methodologies rather convincing.

5.1 Graphical Analysis

We first plot the raw distribution of test statistics as in Brodeur et al. (2016) and Vivalt (2019) to the whole sample. Figure 2 displays an histogram of test statistics for $z \in [0, 10]$ for the entire sample. Bins are 0.1 wide and we superimpose an Epanechnikov kernel.

Three aspects of Figure 2 are worth noting. First, the figure presents an almost

monotonically falling curve with maximum density close to 0. Around one half of test statistics in our sample are null results with large p-values. There is also a small spike of results with z-statistics in a tight range just above 1.96, and an apparent (relatively small) lack of the mass in the range just below that this statistical threshold is potentially due to p-hacking. We formally test for this explanation later.

Second, the distribution of test statistics presented here is remarkably similar to that presented in Brodeur et al. (2016) and Brodeur et al. (2020) for their subsamples of RCTs for the years 2005–2011 and 2015 and 2018, respectively. This finding suggests little has changed over time, despite the apparent increase in awareness of transparency and credibility issues in the research community in more recent years.

Third, the results presented here provide additional evidence that RCTs are less prone to p-hacking and publication bias than other methods of empirical work used in these economics journals such as instrumental variables (Brodeur et al. (2020); Vivalt (2019)).

We now turn to our main research question and investigate whether the distribution of test statistics varies by pre-registration status. We would expect to see less of a bump near the 5% statistical significance threshold if pre-registration does reduce the capacity for researchers to p-hack. Figure 3 presents our main results.[11] In this figure, we decompose our sample based on pre-registration status (i.e., whether the RCT was registered before its trial end date), and plot the distributions for each subsample. Visually, there does not seem to be any discernible change by pre-registration status. This is confirmed by a Kolmogorov–Smirnov test which does not reject the null of equality of distributions at the 10% level.[12]

[11] In Appendix Figures A2 and A3, we use the inverse of the number of tests presented in the same article to weight observations. Re-weighting is potentially problematic in our setting given that the number of test statistics reported might be an outcome of pre-registration.

[12] In Appendix Figures A4 and A5, we deal with rounding issues in our sample. As in Brodeur et al. (2016), a small proportion of coefficients and standard errors are reported with poor precision. For example, we would reconstruct a z-statistic of 2 if the coefficient is 0.020 and the standard error is 0.010. Hence, reconstructed z-statistics are over-abundant for fractions of integers. To deal with this issue, we smooth the distribution by randomly redrawing a value in the interval of potential z-statistics given the reported values and their precision. We follow Brodeur et al. (2016) and use a uniform distribution.

This result may be due to the small amount of p-hacking by RCT practitioners in economics. Another possibility is that pre-registering a RCT is not sufficient as authors might not pre-specify the whole set of hypotheses that will be tested. We explore this explanation in Section 6. A third explanation is that our definition of pre-registration is too loose. In Appendix Figure A1, we replicate Figure 3 but instead code RCTs as pre-registered if the initial registration date is before its trial start date. This definition is quite strict as only 29 articles in our sample are considered pre-registered. The distribution of test statistics presented in Appendix Figure A1 for pre-registered RCTs also exhibits a bump near the 5% statistical significance and appears to have slightly *less* estimates with large p-values.

Overall, our graphical analysis offers little evidence that pre-registration reduces p-hacking.

5.2 Caliper Tests

Caliper tests are well-established tools in research on p-hacking and publication bias, and involve inspection of the prevalence of test statistics within a narrow band either side of arbitrary significance thresholds. The main advantage of this method is that we are able to control for confounders in our estimation. We focus throughout on the 5% and 10% significance thresholds, and show estimates for the 1% threshold in the appendix.

For the 5% threshold:

$$R_{-,h} = [1.96 - h, 1.96], R_{+,h} = [1.96, 1.96 + h] \tag{2}$$

for a bandwidth parameter h, we estimate the following equation:

$$Pr(Significant_{ij} = 1) = \Phi(\alpha + \beta_j + X'_{ij}\delta + \gamma Preregist_{ij}) \tag{3}$$

where $Significant_{ij}$ is a dummy variable for that takes the value 1 if test i in journal j is statistically significant at the 5%-level, zero otherwise. We rely on

probit models and in our main specification report standard errors clustered at the journal article-level.[13] The variable of interest is $Preregist_{ij}$, which represents a dummy variable for whether the test is in an article that has been pre-registered.

The estimates are presented in Table 4 for the 5% statistical significance threshold. (See Appendix Tables A3 and A4 for the other statistical significance thresholds.) In columns 1–3, we restrict the sample to $z \in [1.46, 2.46]$ for the 5% statistical significance. Our sample size for the caliper test using $z \in [1.46, 2.46]$ is 3,870.[14] We also check the robustness of our results to smaller bandwidths in columns 4 ($z \in [1.61, 2.31]$) and 5 ($z \in [1.76, 2.16]$). We find that test statistics in pre-registered RCTs are not less likely to be marginally statistically significant than an estimate in not pre-registered RCTs. The point estimate is is very small (-0.018) and statistically (in)significant at the 51% level.

In columns 2–5, we add controls for the share of authors at top institutions, the share of authors who graduated from a top institution, the share of female authors, an indicator for whether at least one of the authors was an editor of an economics journal at the time of publication, year fixed effects, and reporting a t-statistic, p-value or coefficient and standard error. In column 2, we add dummy variables for Top 5 journals and the AEA journals. In columns 3–5, we instead add journal fixed effects. The point estimates are all negative, very small and statistically insignificant. The results are similar for the other statistical significance thresholds (Appendix Tables A3 and A4).[15]

5.3 Excess Test Statistics Method Proposed by Brodeur et al. (2016)

We now move to a method developed in Brodeur et al. (2016) to quantify the excess (or dearth) of p-values over various ranges by comparing the observed distribution of test statistics for each pre-registration status to a counterfactual distribution

[13] Our results are robust to using logit models. See Appendix Table A2.

[14] We estimate that we have well over 95% power to detect an effect of 0.025. See Appendix Figure A6.

[15] Our results are also robust to weighting. In Appendix Table A5, we use the inverse of the number of tests presented in the same article to weight observations.

that we would expect to emerge absent publication bias. We follow Brodeur et al. (2016) by assuming that the observed test statistic distribution above $z = 5$ should be free of p-hacking or publication bias. Given there is little assumed distortion in this tail of the observed distribution, we calibrate (via non-centrality parameter and degrees of freedom)[16] a counterfactual non-central t-distribution that has a near-identical tail. We then produce a separate non-central t-distribution for each of two subgroups of articles– (i) without pre-registration and (ii) pre-registered– that closely fits the observed distribution. in the range $z > 5$ by calibrating the degrees of freedom and non-centrality parameter.

Refining Brodeur et al. (2020), we proceed as follows. For 0 to 5 degrees of freedom, we calculate the non-centrality parameter that minimizes the difference in the $z > 5$ area between the observed distribution and the expected distribution. We then choose the "best" of the optimized t-distributions. In this manner we explore the entire region of $0 < df < 5$ and $0 < np < 5$.

In Figure 4, we present the results for pre-registered (right panel) and non-pre-registered (left panel) RCTs for the following regions: $[0 < z < 1.65)$, $[1.65 < z < 1.96)$, $[1.96 < z < 2.58)$, $[2.58 < z < 5)$ and $[5, \infty)$. These figures illustrate both the observed distribution of test statistics as a solid line–which corresponds directly to the kernel density in Figure 2– and the counterfactual non-central t-distribution in dashes.

Overall, this method suggests that the excess of z-statistics above the 5% significance thresholds is almost identical for both subgroups of RCTs. In the statistically insignificant region of $0 < z < 1.645$, the observed distribution is "missing" 28% and 27% of the total mass for both pre-registered and not pre-registered RCTs. Most of these "missing" test statistics can be found above the 5% statistical significance threshold where there are 10% more than expected and in the $[2.58 < z < 5)$ interval where there are about 8% more than expected.

[16]Degrees of freedom are optimized in steps of 1. The non-centrality parameter of the t-distribution is positive and real valued, optimized in steps of 0.01.

5.4 Tests for p-Hacking Proposed by Elliott et al. (2022)

In this subsection, we rely on Elliott et al. (2022)'s tests to detect p-hacking. Elliott et al. (2022) derive testable restrictions for test statistics resulting in tests against a null hypothesis of no p-hacking. We report histograms of p-values (p-curves) which are truncated above 0.15. The figures contain two types of tests; those based on the non-increasingness of the p-curve and those testing for discontinuities. We illustrate these tests and the p-curves for our two subgroups in Figure 5 and describe the results below.

P-curves should be non-increasing under very general conditions following Theorem 1 in Elliott et al. (2022) including regularity of the underlying test statistics' cumulative distribution function and a restriction of how the power function changes as an examined critical value changes. We discuss the tests embedded in Figure 5. First the binomial test. We follow and use the code of Elliott et al. (2022) to split [0.04,0.05] into two subintervals [0.04,0.045] and (0.04,0.045]. Under the null of no p-hacking, the fraction of p-values in (0.045,0.05] should be smaller than or equal to 0.5. For both non-pre-registered and pre-registered articles the p-value is significant at the 1% level ($p = 0.000$ and $p = 0.001$ respectively.) Second, Fisher's test categorically compares the significant to not statistically significant test statistics (both p-values are effectively 1, which is exactly what Elliott et al. (2022) find in their applications as well). Third, CS1 is an application of the conditional chi-squared test of Cox and Shi (2022). For both non-pre-registered and pre-registered samples the p-value is once again less than 0.01. The same is true for the 'more powerful' CS2B, introduced by Elliott et al. (2022) and is another histogram based test designed against its 2-monotonicity and places bounds on the p-curve and its first two derivatives. Fifth, and our first major difference between the samples, a test whose null hypothesis is that the p-value distribution (not the p-curve) is concave. Applying this Least Concave Majorant (LCM) test (Beare and Moon 2015), we find it is statistically significant for the non-pre-registered sample ($p = 0.010$) and not for the pre-registered sample ($p = 0.948$) indicating a deviation from expected in

the non-pre-registered articles only.

Finally, we also provide a discontinuity test (an application of the density discontinuity test from Cattaneo et al. (2020)) which rejects the null hypothesis of no discontinuity in the non-pre-registered sample ($p = 0.048$) and fails to reject it in the pre-registered sample ($p = 0.188$).

In summary, five of the tests included in Elliott et al. (2022)'s reject their null hypothesis for the non-pre-registered sample while only 3 joint tests for publication bias or p-hacking reject their null hypothesis for the pre-registered sample. While these tests do not directly compare p-hacking and publication bias rates across samples, they do suggest that both pre-registered and non-pre-registered RCTs suffer, to some extent and rather similarly, from these biases.

5.5 Identification of Publication Bias Following Andrews and Kasy (2019)

Last, we investigate whether pre-registration decreases publication bias (i.e., statistical significance of a result determines the probability of publication.). While pre-registration may not have an impact on p-hacking, it is still plausible that it increases the likelihood of null results being published in leading economics journals.

We rely on Andrews and Kasy (2019)'s method for identifying the conditional probability of publication as a function of a study's results. The primary estimated parameter provided by Andrews and Kasy (2019) is the relative publication probability of a statistically significant result compared to a statistically *in*significant result. The results are reported in Table 5. We also present the generalized t distribution parameters the model fits for the underlying effect distribution.

In our complete sample, a statistically significant result is estimated to be 1.65 times *more* likely to be published than a statistically insignificant result. Conversely, a statistically insignificant result is 0.60 times as likely as a statistically significant one to be published, indicating a relatively modest degree of publication bias.[17] The estimated parameter for pre-registered RCTs is 1.65, which is slightly smaller

[17]Using the same method, Brodeur et al. (2020) find an estimated relative publication probability of 4.72 times for studies using instrumental variables.

in comparison to 1.67 for non-pre-registered RCTs.

To sum up, we find limited evidence that pre-registration significantly reduce the extent of p-hacking for RCTs in economics. The estimated effect on publication bias is also small.

6 Mechanisms

6.1 Reporting

Our analysis of pre-registration uncovered a very small, if any, reduction in p-hacking and publication bias. One interesting feature of our data is that we collected all coefficients of interest for all RCTs. On average, we collected about 51 test statistics per article. We investigate now whether authors of pre-registered RCTs report a larger number of test statistics. One advantage of pre-registration could be that it disciplines authors to be more transparent and report results for all econometric specifications and outcome variables that were pre-specified.

We provide some evidence that pre-registration leads researchers to report more results. On the one hand, we collected on average 58.94 (std. dev. 49.40) test statistics for pre-registered RCTs in comparison to 48.05 (std. dev. 40.89) for not pre-registered RCTs. The difference is statistically significant at conventional levels. On the other hand, we find that the test statistics collected are reported in 2.9 tables on average for both pre-registered and non-pre-registered RCTs.

Overall, these results provide suggestive evidence that while pre-registration does not meaningfully reduce p-hacking it does increase reporting of test statistics. This finding is consistent with low levels of p-hacking among RCT users in economics, and pre-registration simply increasing the number of results and specifications reported.

6.2 Pre-analysis Plans

So far our findings indicate that pre-registration in itself does not appear to play an important role in reducing p-hacking in our sample.

As already noted, a possible explanation for this is that pre-registration in itself

does not typically require a tight or comprehensive statement or hypotheses that will be tests and how data will be handled. Such details are, however, part of a typical PAP. To explore the additional discipline imposed by pre-registration of a PAP we exploit that the AEA RCT repository provides the option - but does not require - for authors to pre-register a fully-fledged "Analysis Plan". Hence not all pre-registrations contain fully written PAPs, though pre-registrations will typically contain a subset of line items that could be used in a PAP (e.g., listing "Primary Outcomes", the "Randomization Method", "Was the treatment clustered?", etc.).

We further test whether the extent of publication bias and p-hacking differ for subsamples such as pre-registered RCTs with and without a complete PAP. Our baseline analysis did not differentiate between these two categories and considers all RCTs that were pre-registered as 'treated'.

We code a complete PAP as a pre-registration made before the list trial end date that contained some form of a write-up document. In the AEA RCT registry, this is the optional "Analysis Plan" write-up attachment. If a pre-registration made before the end trial date did not contain such a write-up such a registration would not be counted having a PAP. Also if the registration was made after the end trial date, it would not be counted as a PAP nor a pre-registration.

Figure 6 illustrates the distribution of test statistics for pre-registered RCTs for those containing a PAP (right panel) and those without a PAP (left panel), respectively.[18] Both curves are monotonically falling with a bump around the 5% significance threshold. But, of note, visually the bump is more pronounced for the curve without a PAP and the curve for PAP contains more tests with large p-values.

Our caliper tests confirm these patterns. In Table 6, we replicate the structure and specifications of Table 4, replacing the dummy variable pre-registration by a dummy for whether the test statistic is in an RCT containing a PAP. We restrict the sample to pre-registered RCTs. We have 1,164 observations for our baseline window at the 5% significance level. Our estimates are all negative and suggest

[18] See Appendix Figure A7 for the derounded distributions.

that the proportion of test statistics that are marginally significant in articles with a complete PAP is about 10 percentage points lower than in pre-registered RCTs without a PAP. Four and out five estimates are statistically significant at the 5% level, including all our specifications with journal fixed effects.

We also find that the extent of publication bias is lower for pre-registered RCTs with a PAP in comparison to those without a PAP (see Table 5). Other things equal, a statistically significant result is estimated to be 1.38 times *more* likely to be published than a statistically insignificant result for pre-registered RCTs with a PAP against 2.09 times for those without a PAP.

6.3 Discussion of Statistical Power

We further document the relationship between p-hacking, publication bias and the completeness of a pre-registration by creating a variable, following Ofosu and Posner (2021), for whether the PAP included a power analysis.[19] Figure 7 shows the distribution of test statistics for pre-registered RCTs for those including a discussion of statistical power (right panel) and those without such a discussion (left panel), respectively.[20] Visually, we find that the distributions are very different (a two-sided Kolmogorov–Smirnov test also rejects the null of equality of distributions with $p < 0.000$), with a bump around the 5% significance threshold for articles that do not include an explicit discussion of statistical power and a monotonically falling curve without a bump for RCTs including such a discussion.

[19] This exercise was not specified in our own PAP and so should be regarded as exploratory. Ofosu and Posner (2021) also code other elements of the 'quality' of PAPs, such as whether the authors specified a clear hypothesis; specified the primary dependent and independent variables sufficiently clearly so as to prevent post-hoc adjustments; and spelled out the precise statistical model to be tested including functional forms and estimator. We decided against coding pre-registered RCTs in our sample along these dimensions because of the subjectivity in coding some of these variables. Moreover, virtually all pre-registered RCTs in our sample described the main variables as this is an obligatory field in the AEA RCT registry.

[20] See Appendix Figure A8 for the derounded distributions.

7 Conclusion

The credibility of the published body of empirical research in the social sciences has come under increasing challenge in recent years, with p-hacking and publication bias understood to pose an important threat to that credibility. Among economists several new 'open science' practices have been promoted as ways forward to mitigate such credibility concerns. Favored among these have been pre-registration of studies and pre-analysis plans. The aim in this paper to provide a systematic investigation of the efficacy of these practices by comparing, using a variety of state-of-the-art methods, the patterns of published test statistics observed in RCTs that varied in pre-registration status.

Our main results are both optimistic and pessimistic.

First, published RCT-based research in economics appears to be characterized by comparatively little p-hacking and publication bias. Second, pre-registration *in itself* does not appear to be an important driver of that - there is little discernible difference between the pattern of test statistics found in pre-registered studies and those from their non-pre-registered counterparts. Third, however, pre-registration that includes a fully-fledged PAP *does* appear to significantly reduce p-hacking, a result which remains robust across a variety of different methods and specifications.

These results point to a potentially important ameliorating effect of pre-registration combined with pre-analysis plans that presents an argument in favor of wider adoption of more detailed pre-registration practices, to include PAPs. To be useful in bolstering credibility registries may want to adopt correspondingly more stringent pre-registration requirements. The AEA RCT registry for instance does not currently require the inclusion of written PAPs and within registrations certain line items which also appear to reduce p-hacking, like the inclusion of power analyses, remain optional for researchers to complete.

Our analysis faces limitations arising from the observational nature of our data. For example, the choice of whether or not to pre-register an RCT is plausibly

not made randomly, and we aim to document the characteristics of researchers adopting this research practice. Another limitation is that we focus on leading economic journals. It thus remains unclear if the distribution of test statistics for pre-registered and non-pre-registered RCTs differ at lower ranked journals, outside of the top 15 which are the source of our sample.

Despite these shortcomings, we believe that our data set and analysis provides the best available setup to investigate the relationship between p-hacking, publication bias and pre-registration. Not only do we analyze the universe of test statistics for a set of highly-regarded and reputable economics journals, but our study window also straddles period over which the prevalence of pre-registration has grown substantially, from a rarity to comparatively commonplace. By using multiple, state of the art methods - each with their own strengths and weaknesses, and subject to specific critiques - our investigation provides a more comprehensive picture than would a study relying on a single approach.

Whether the benefits in terms of enhanced credibility identified here justify the costs of mandating PAPs more widely is left for future research. So too the efficacy of other open science practices, such as registered reports and changing significance thresholds (Benjamin et al. (2018) and Ludwig et al. (2019)).

References

Andrews, I. and Kasy, M.: 2019, Identification of and Correction for Publication Bias, *American Economic Review* **109**(8), 2766–94.

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L. F., Olken, B. A. and Sautmann, A.: 2020, In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics. NBER Working Paper 26993.

Beare, B. K. and Moon, J.-M.: 2015, Nonparametric tests of density ratio ordering, *Econometric Theory* **31**(3), 471–492.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C. et al.: 2018, Redefine Statistical Significance, *Nature Human Behaviour* **2**(1), 6–10.

Blanco-Perez, C. and Brodeur, A.: 2020, Publication Bias and Editorial Statement on Negative Findings, *Economic Journal* **130**(629), 1226–1247.

Brandon, A. and List, J. A.: 2015, Markets for Replication, *Proceedings of the National Academy of Sciences* **112**(50), 15267–15268.

Brodeur, A., Cook, N. and Heyes, A.: 2020, Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics, *American Economic Review* **110**(11), 3634–60.

Brodeur, A., Cook, N. and Neisser, C.: 2022, P-Hacking, Data Type and Availability of Replication Material. University of Ottawa mimeo.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y.: 2016, Star Wars: The Empirics Strike Back, *American Economic Journal: Applied Economics* **8**(1), 1–32.

Burlig, F.: 2018, Improving Transparency in Observational Social Science Research: A Pre-Analysis Plan Approach, *Economics Letters* **168**, 56–60.

Butera, L., Grossman, P. J., Houser, D., List, J. A. and Villeval, M.-C.: 2020, A New Mechanism to Alleviate the Crises of Confidence in Science-with an Application to the Public Goods Game. NBER Working Paper 26801.

Camerer, C. F., Dreber, A. and Johannesson, M.: 2019, Replication and Other Practices for Improving Scientific Quality in Experimental Economics, *Handbook of Research Methods and Applications in Experimental Economics* .

Casey, K., Glennerster, R. and Miguel, E.: 2012, Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan, *Quarterly Journal of Economics* **127**(4).

Cattaneo, M. D., Jansson, M. and Ma, X.: 2020, Simple local polynomial density estimators, *Journal of the American Statistical Association* **115**(531), 1449–1455.

Christensen, G., Dafoe, A., Miguel, E., Moore, D. A. and Rose, A. K.: 2019, A Study of the Impact of Data Sharing on Article Citations Using Journal Policies as a Natural Experiment, *PLoS One* **14**(12), e0225883.

Christensen, G. and Miguel, E.: 2018, Transparency, Reproducibility, and the Credibility of Economics Research, *Journal of Economic Literature* **56**(3), 920–80.

Christensen, G., Wang, Z., Levy Paluck, E., Swanson, N., Birke, D., Miguel, E. and Littman, R.: 2020, Open Science Practices Are on the Rise: The State of Social Science (3S) Survey. https://osf.io/preprints/metaarxiv/5rksu/.

Coffman, L. C. and Niederle, M.: 2015, Pre-analysis plans have limited upside, especially where replications are feasible, *Journal of Economic Perspectives* **29**(3), 81–98.

Cox, G. and Shi, X.: 2022, Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models, *Review of Economic Studies* .

Doucouliagos, C. and Stanley, T. D.: 2013, Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity, *Journal of Economic Surveys* **27**(2), 316–339.

Drazen, A., Dreber, A., Ozbay, E. Y. and Snowberg, E.: 2021, Journal-Based Replication of Experiments: An Application to "Being Chosen to Lead", *Journal of Public Economics* **202**, 104482.

Elliott, G., Kudrin, N. and Wüthrich, K.: 2022, Detecting p-Hacking, *Econometrica* **90**(2), 887–906.

Fang, Albert, G. G. and Humphreys, M.: 2015, Does Registration Reduce Publication Bias? No Evidence from Medical Sciences. Working Paper.

Franco, A., Malhotra, N. and Simonovits, G.: 2016, Underreporting in Psychology Experiments: Evidence from a Study Registry, *Social Psychological and Personality Science* **7**(1), 8–12.

Furukawa, C.: 2019, Publication Bias under Aggregation Frictions: From Communication Model to New Correction Method. MIT Mimeo.

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

Havránek, T.: 2015, Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting, *Journal of the European Economic Association* **13**(6), 1180–1204.

Havránek, T. and Sokolova, A.: 2020, Do Consumers Really Follow a Rule of Thumb? Three Thousand Estimates from 144 Studies Say "Probably Not", *Review of Economic Dynamics* **35**, 97–122.

Ludwig, J., Mullainathan, S. and Spiess, J.: 2019, Augmenting Pre-Analysis Plans with Machine Learning, *AEA Papers and Proceedings*, Vol. 109, pp. 71–76.

Miguel, E.: 2021, Evidence on Research Transparency in Economics, *Journal of Economic Perspectives* **35**(3), 193–214.

Ofosu, G. K. and Posner, D. N.: 2020, Do Pre-Analysis Plans Hamper Publication?, *AEA Papers and Proceedings*, Vol. 110, pp. 70–74.

Ofosu, G. K. and Posner, D. N.: 2021, Pre-Analysis Plans: An Early Stocktaking, *Perspectives on Politics* pp. 1–17.

Olken, B. A.: 2015, Promises and Perils of Pre-Analysis Plans, *Journal of Economic Perspectives* **29**(3), 61–80.

Ravallion, M.: 2020, Should the Randomistas (Continue to) Rule? NBER Working Paper 27554.

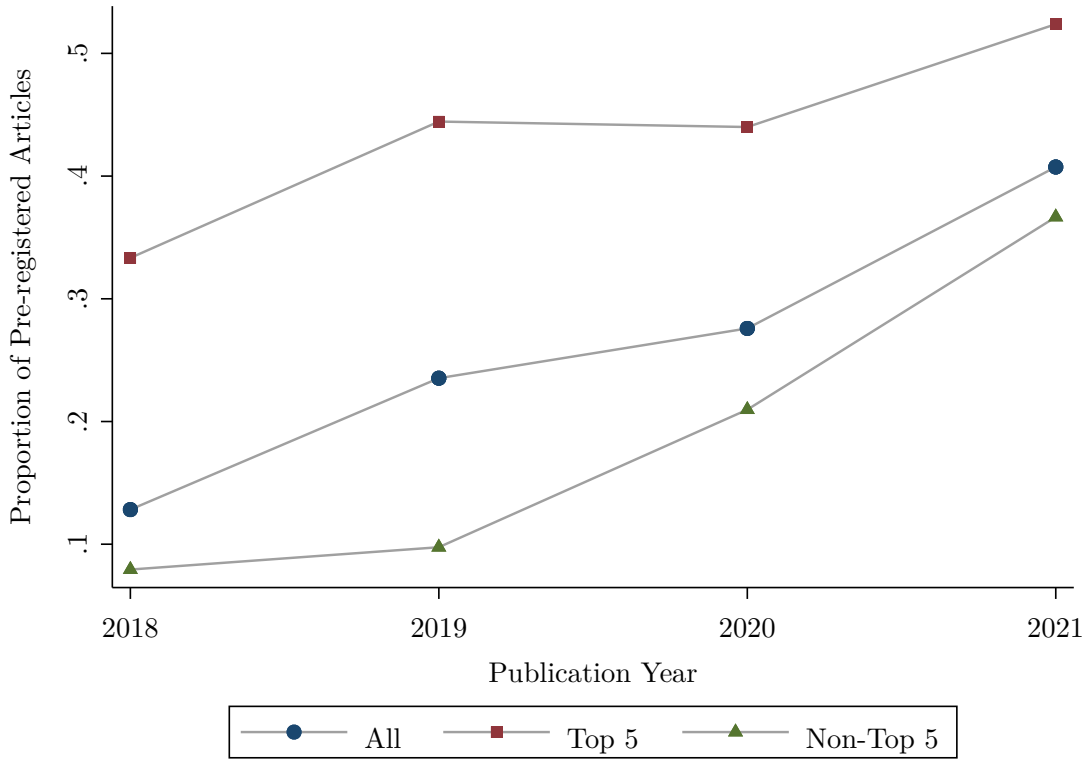
Scheel, A. M., Schijen, M. R. and Lakens, D.: 2021, An Excess of Positive Results: Comparing the Standard Psychology Literature with Registered Reports, *Advances in Methods and Practices in Psychological Science* **4**(2), 25152459211007467.

Stanley, T. D. and Doucouliagos, H.: 2014, Meta-regression Approximations to Reduce Publication Selection Bias, *Research Synthesis Methods* **5**(1), 60–78.

Vivalt, E.: 2019, Specification Searching and Significance Inflation Across Time, Methods and Disciplines, *Oxford Bulletin of Economics and Statistics* **81**(4), 797–816.

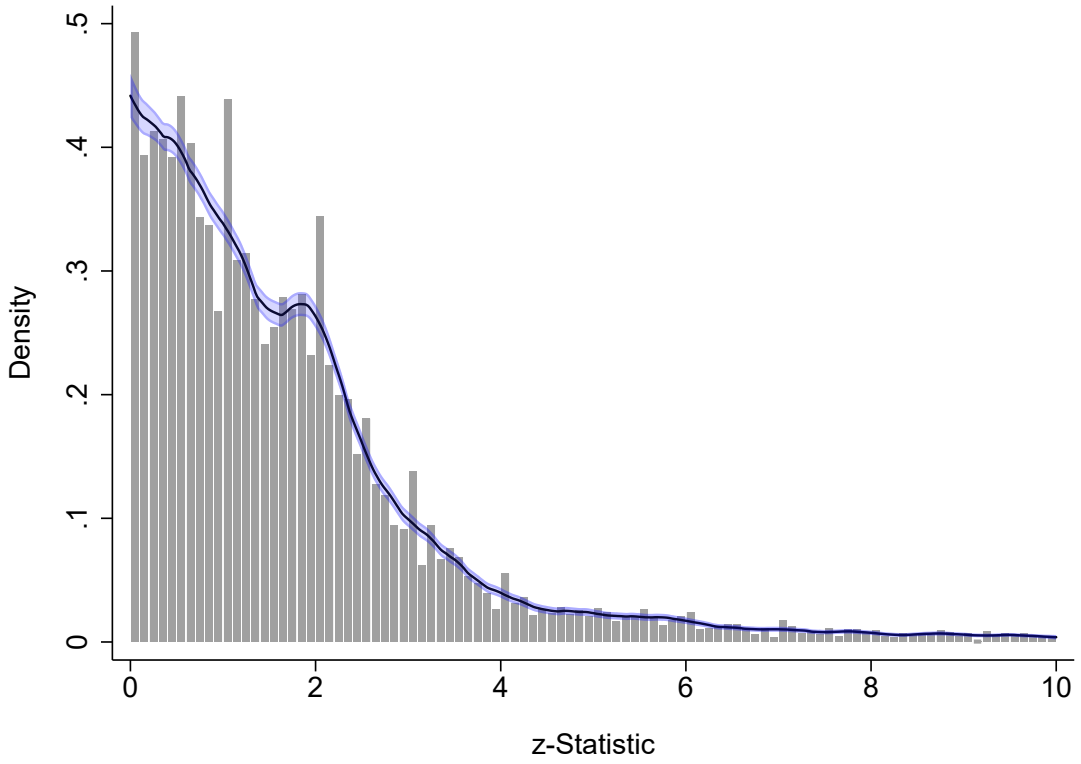
8 Figures

Figure 1: Pre-Registration Rates Over Time



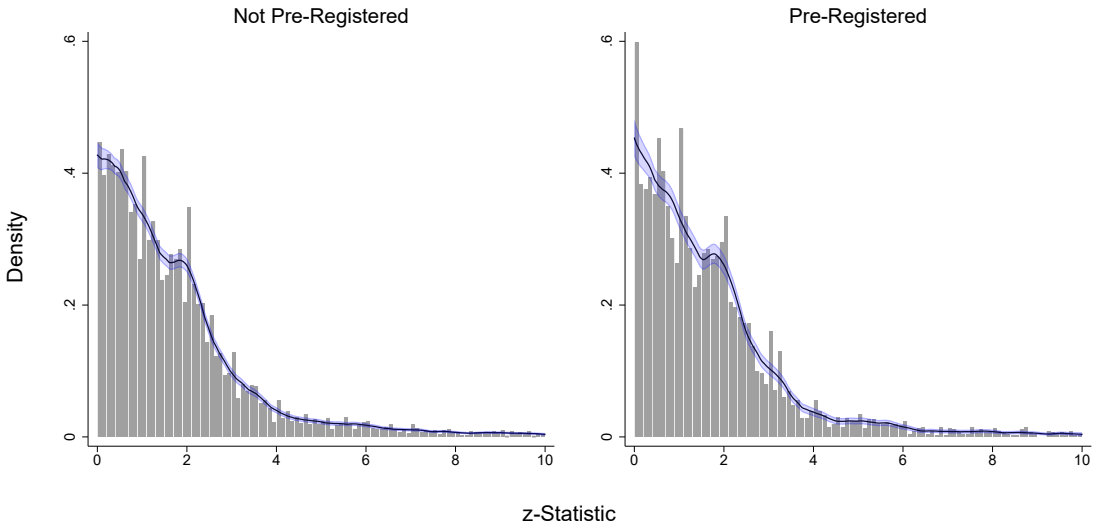
Notes: This figure displays the percentage of pre-registered RCT articles in full our sample and for Top 5 and non-Top 5 journals. We define a pre-registered RCT as a study that was registered before its trial end date listed in a registry. Studies that were registered after the trial end date are counted as non-pre-registered. The Top 5 journals in economics are the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*.

Figure 2: Test Statistics Distribution



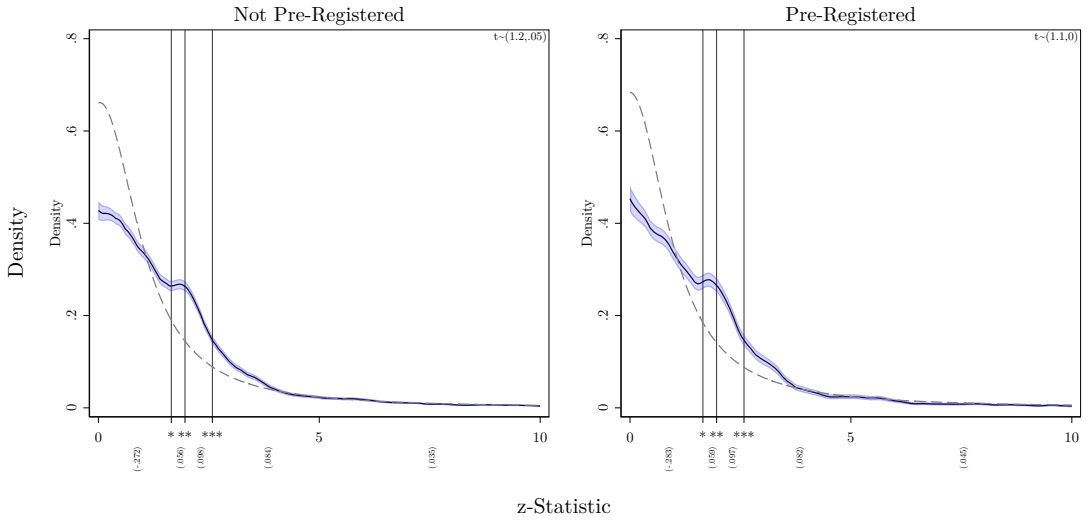
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials from 2018–2021. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

Figure 3: Test Statistics Distribution by Pre-Registration Status



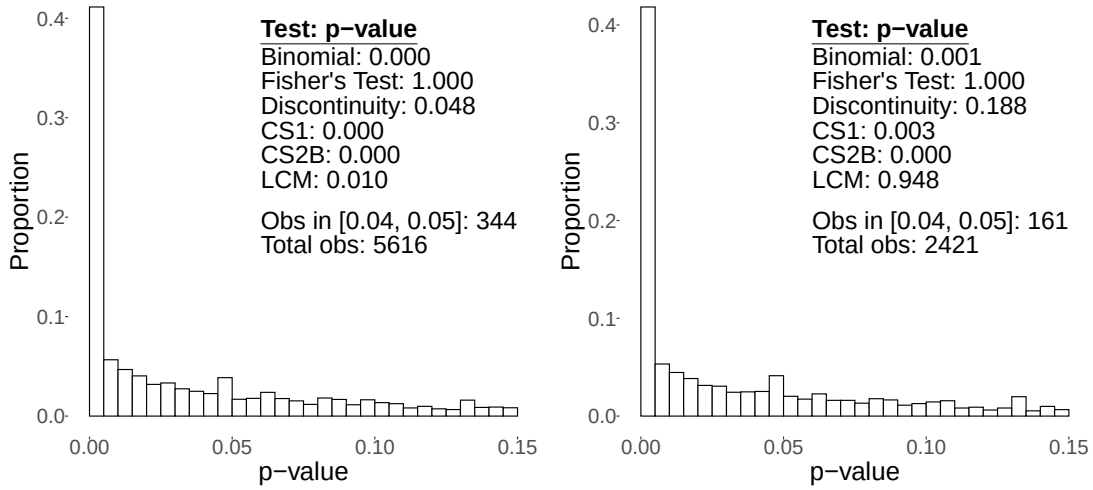
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials from 2018–2021 by pre-registration status. We define a pre-registered RCT as a study that was registered before its trial end date listed in a registry. Studies that were registered after the trial end date are counted as non-pre-registered. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

Figure 4: Excess Test Statistics by Pre-Registration Status



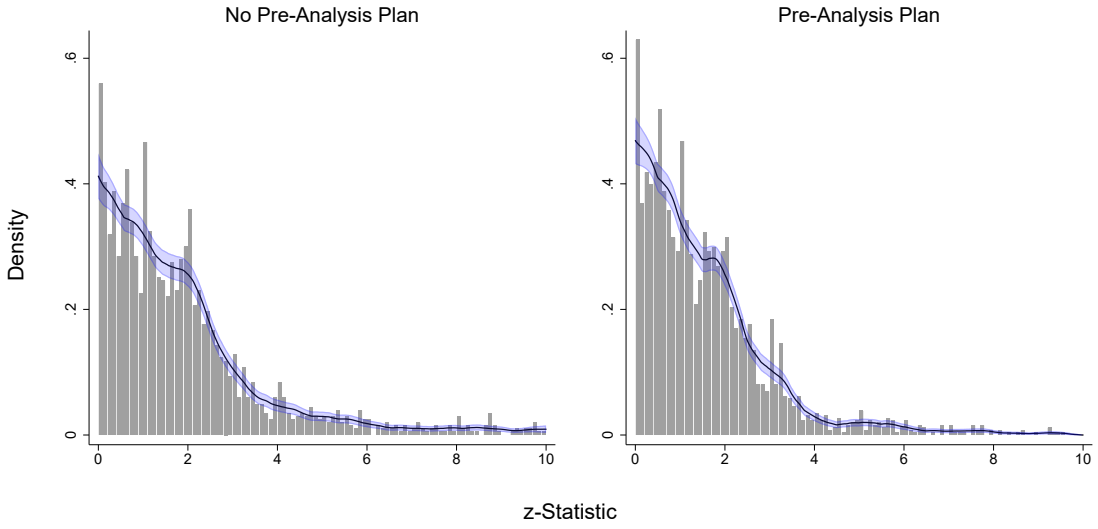
Notes: This figure presents the observed distribution of test statistics by pre-registration status. We also present status-specific counterfactual distributions we expect to observe in the absence of publication bias or p-hacking (see Brodeur et al. (2016) and Brodeur et al. (2020)). Below the horizontal axis we include the difference in mass between statistical significance thresholds.

Figure 5: Application of Elliott et al. (2022) by Pre-Registration Status



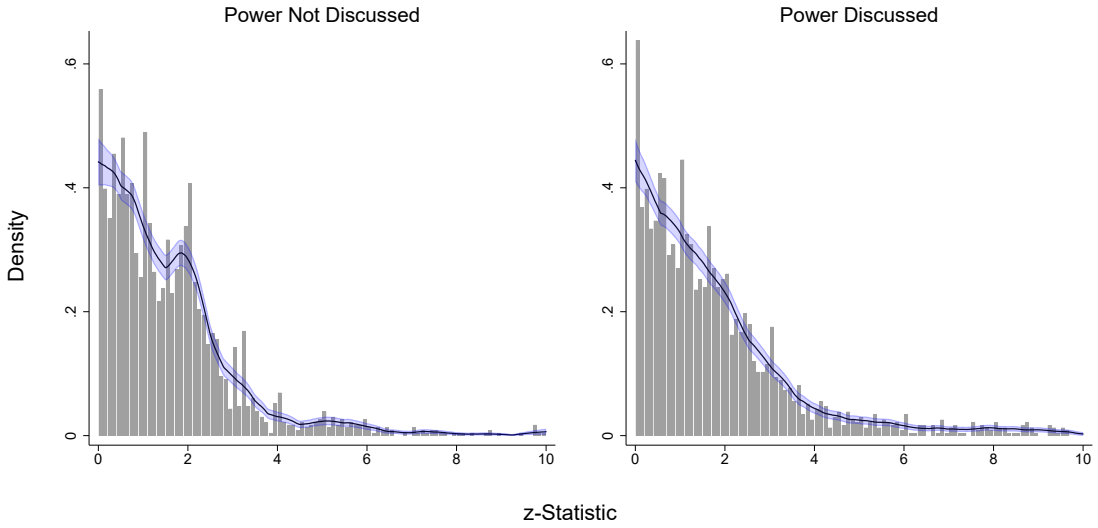
Notes: Each panel is a direct application of Elliott et al. (2022)'s p-hacking test battery to a sub sample. The left panel shows the p-curve not pre-registered RCTs. The right panel illustrates the p-curve for pre-registered RCTs. Section 5.4 discusses the p-curve histograms and included tests in detail.

Figure 6: Test Statistics Distribution for Pre-Registered RCTs by a Presence of Pre-Analysis Plan



Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials that were pre-registered from 2018–2021 by presence of a pre-analysis plan. We define a pre-registered RCT as a study that was registered before its trial end date listed in a registry. Studies that were registered after the trial end date are counted as non-pre-registered. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

Figure 7: Test Statistics Distribution for Pre-Registered RCTs by Presence of Power Analysis



Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials that were pre-registered from 2018–2021 by power analysis status. We define a pre-registered RCT as a study that was registered before its trial end date listed in a registry. Studies that were registered after the trial end date are counted as non-pre-registered. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

9 Tables

Table 1: Summary Statistics

	Mean	Std. Dev.	Min	Max
	(1)	(2)	(3)	(4)
Pre-Registered	0.30	0.46	0	1
Year	2019	1.08	2018	2021
Solo Authored	0.07	0.26	0	1
Number Authors	3.29	1.43	1	10
% Female Authors	0.32	0.33	0	1
Top 5 Journals	0.30	0.46	0	1
AEA Journals	0.30	0.46	0	1
Average Experience	11.70	5.01	1	33.33
Share Top Institutions	0.25	0.33	0	1
Share Top (PhD) Institutions	0.45	0.35	0	1
Editor Present	0.65	0.48	0	1

Notes: This table provides summary statistics; mean (standard error). The unit of observation is a test statistic.

Table 2: Summary Statistics by Pre-Registration Status

	All	Not Pre-Registered	Pre-registered
	(1)	(2)	(3)
Year	2019	2019	2020
	(1.08)	(1.06)	(1.04)
Solo Authored	0.07	0.10	0.03
	(0.26)	(0.29)	(0.16)
Number Authors	3.29	3.19	3.53
	(1.43)	(1.46)	(1.30)
% Female Authors	0.32	0.35	0.25
	(0.33)	(0.33)	(0.31)
Top 5 Journals	0.30	0.20	0.54
	(0.46)	(0.40)	(0.50)
AEA Journals	0.30	0.27	0.37
	(0.46)	(0.44)	(0.48)
Average Experience	11.70	11.75	11.58
	(5.01)	(5.08)	(4.84)
Share Top Institutions	0.25	0.24	0.28
	(0.33)	(0.33)	(0.32)
Share Top (PhD) Institutions	0.45	0.43	0.49
	(0.35)	(0.35)	(0.34)
Editor Present	0.65	0.68	0.60
	(0.48)	(0.47)	(0.49)

Notes: This table provides summary statistics. Column 1 includes all RCTs. Column 2 restricts the sample to RCTs that were pre-registered. Column 3 restricts the sample to RCTs that were not pre-registered. The unit of observation is a test statistic.

Table 3: Prediction of Pre-Registration Use

	Pre-Registration					
	(1)	(2)	(3)	(4)	(5)	(6)
2019	0.058	0.008	0.029	0.049	-0.030	0.010
	(0.090)	(0.089)	(0.089)	(0.128)	(0.125)	(0.121)
2020	0.134	0.099	0.078	0.125	0.062	0.043
	(0.081)	(0.082)	(0.080)	(0.119)	(0.111)	(0.101)
2021	0.319	0.264	0.261	0.370	0.263	0.261
	(0.093)	(0.087)	(0.084)	(0.129)	(0.112)	(0.108)
Solo Authored	-0.452	-0.435	-0.416	-0.599	-0.599	-0.549
	(0.111)	(0.107)	(0.101)	(0.147)	(0.142)	(0.125)
% Female Authors	-0.238	-0.198	-0.234	-0.372	-0.276	-0.305
	(0.099)	(0.089)	(0.083)	(0.142)	(0.116)	(0.100)
Avg. Experience	0.009	0.004	0.001	0.037	0.018	0.002
	(0.021)	(0.019)	(0.019)	(0.034)	(0.026)	(0.024)
Avg. Experience^2	-0.000	-0.000	0.000	-0.001	-0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.024)
Top Institution	0.131	0.010	0.014	0.159	-0.010	-0.024
	(0.101)	(0.093)	(0.095)	(0.142)	(0.122)	(0.115)
PhD Top Institution	0.179	0.132	0.124	0.197	0.136	0.098
	(0.100)	(0.092)	(0.088)	(0.152)	(0.134)	(0.124)
Editor present	-0.142	-0.152	-0.151	-0.107	-0.139	-0.101
	(0.073)	(0.065)	(0.062)	(0.115)	(0.094)	(0.085)
Top 5		0.252			0.310	
		(0.058)			(0.066)	
AEA Journals		0.016			-0.028	
		(0.065)			(0.080)	
Journal FE			Yes			Yes
Observations	16,064	16,064	15,911	16,064	16,064	15,911
Pseudo R2	0.133	0.199	0.257	0.173	0.262	0.341
Weights				Article	Article	Article

Notes: This table reports marginal effects from probit regressions. The dependent variable is a dummy for pre-registration. Robust standard errors are in parentheses, clustered by article. Observations are unweighted in columns 1–3. In columns 4–6, we use the inverse of the number of tests presented in the same article to weight observations.

Table 4: Caliper Test, Statistically Significant at the 5 Percent Level

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	-0.018	-0.026	-0.023	-0.013	-0.021
	(0.027)	(0.026)	(0.027)	(0.029)	(0.035)
Controls					
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Articles' Charact.		Y	Y	Y	Y
AEA & Top 5		Y			
Journal FE			Y	Y	Y
Observations	3,870	3,779	3,779	2,789	1,671
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]

Notes: This table reports marginal effects from probit regressions. The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table 5: An Application of Andrews and Kasy (2019)'s Method

Sample	Sig. Rel. Pub. Prob.	[0,1.96]	Location	Scale	Degrees of Freedom
Full	1.653	0.605	0.013	0.014	1.716
Not Pre-Registered	1.667	0.600	0.015	0.015	1.794
Pre-Registered	1.647	0.607	0.010	0.009	1.568
No Pre-Analysis Plan	2.088	0.479	0.006	0.004	1.464
Pre-analysis Plan	1.377	0.726	0.014	0.014	1.707
No Power Discussion	1.767	0.566	0.001	0.000	1.683
Power Discussion	1.379	0.725	0.028	0.024	1.526
Gated	1.812	0.552	0.008	0.007	1.515
Ungated	1.272	0.786	0.021	0.023	1.709

Notes: This table presents estimates of the relative publication probability of a statistically significant result. For example, in our entire sample, a z-statistic greater than 1.96 is 1.65 times more likely to be published than a statistically insignificant result. The estimated model uses a non-central t-distribution, whose parameters are reported in the following columns.

Table 6: Caliper Test, Statistically Significant at the 5 Percent Level: Pre-Analysis Plan

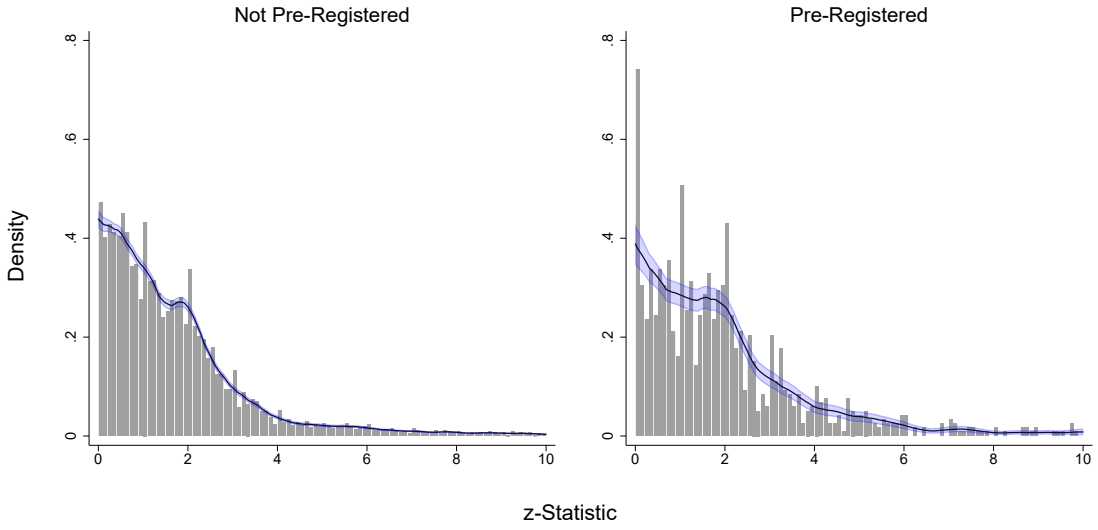
	(1)	(2)	(3)	(4)	(5)
Pre-Analysis Plan	-0.063	-0.108	-0.117	-0.136	-0.113
	(0.046)	(0.044)	(0.043)	(0.049)	(0.051)
Controls					
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Articles' Charact.		Y	Y	Y	Y
AEA & Top 5		Y			
Journal FE			Y	Y	Y
Observations	1,164	1,135	1,131	835	509
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]

Notes: This table reports marginal effects from probit regressions. The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

10 ONLINE APPENDIX

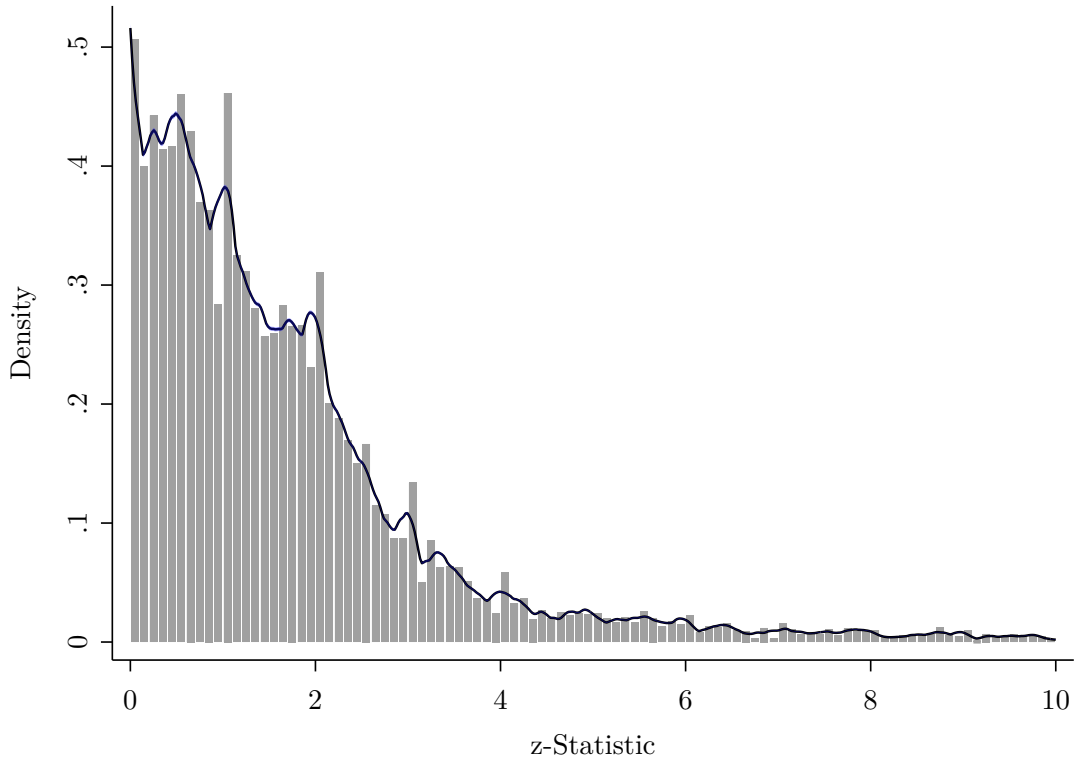
11 Appendix Figures

Figure A1: Robustness Check: Test Statistics Distribution by Pre-Registration Status



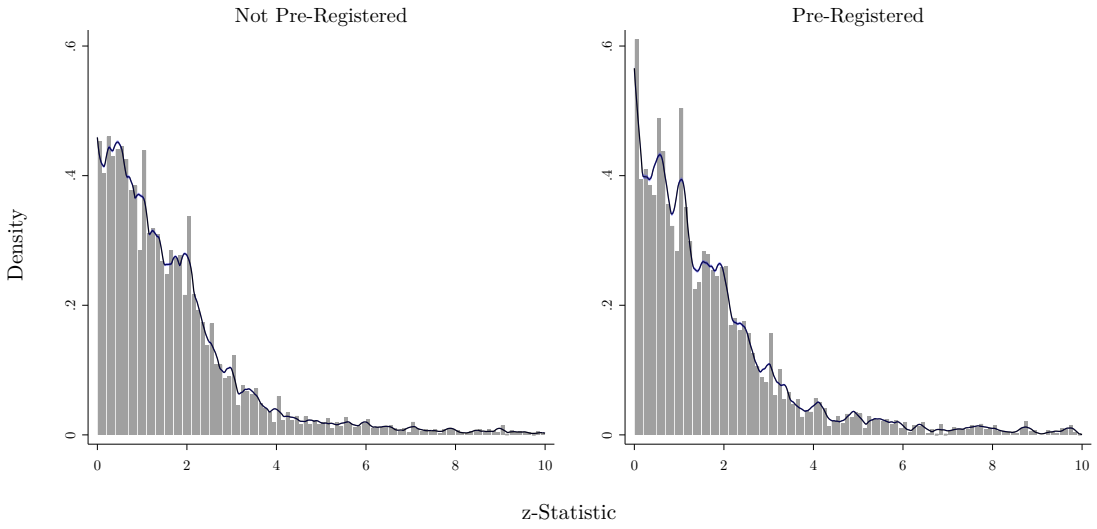
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials from 2018–2021 by pre-registration status. We define a pre-registered RCT as a study that was registered before its trial *start* date listed in a registry. This alternate definition codes as not pre-registered those RCTs that are pre-registered after the trial start date. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

Figure A2: Test Statistics Distribution: Article Weights



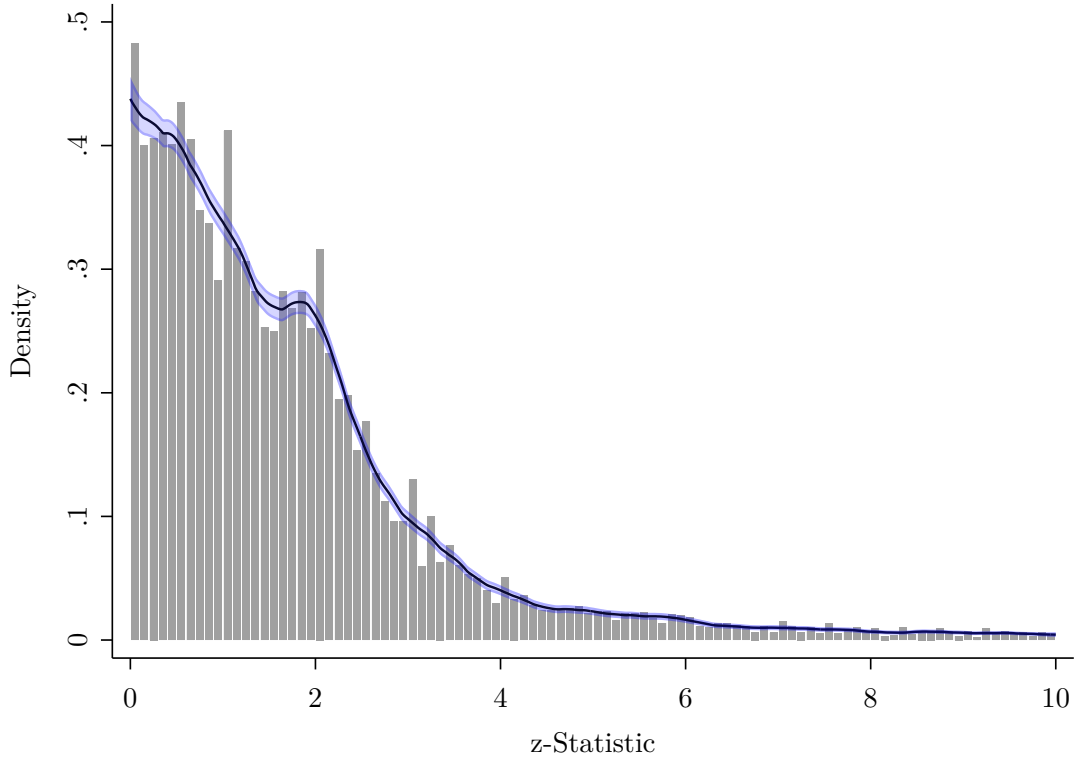
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials from 2018–2021. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A3: Test Statistics Distribution by Pre-Registration Status: Article Weights



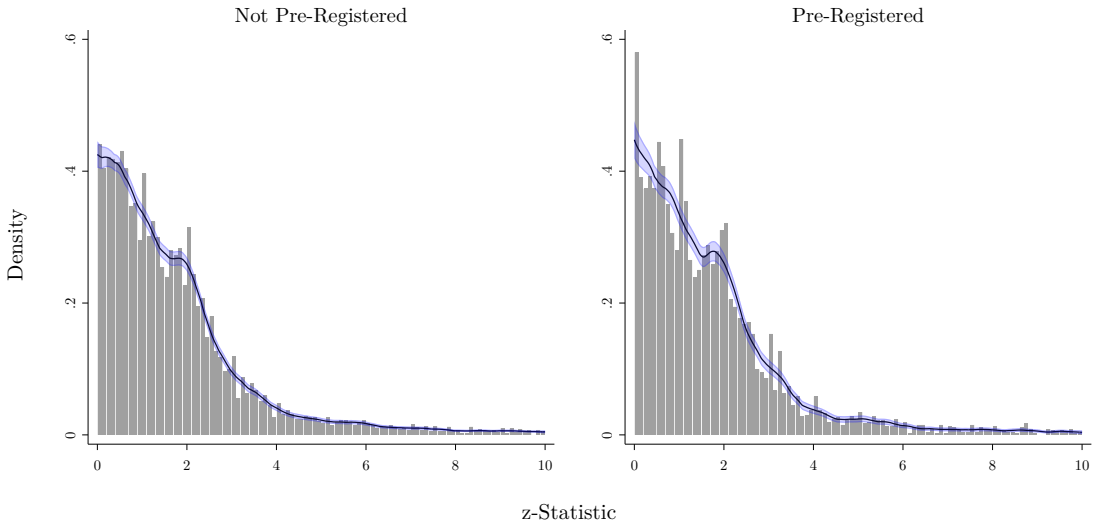
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials from 2018–2021 by pre-registration status. We define a pre-registered RCT as a study that was registered before its trial end date listed in a registry. Studies that were registered after the trial end date are counted as non-pre-registered. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A4: Test Statistics Distribution: Derounding



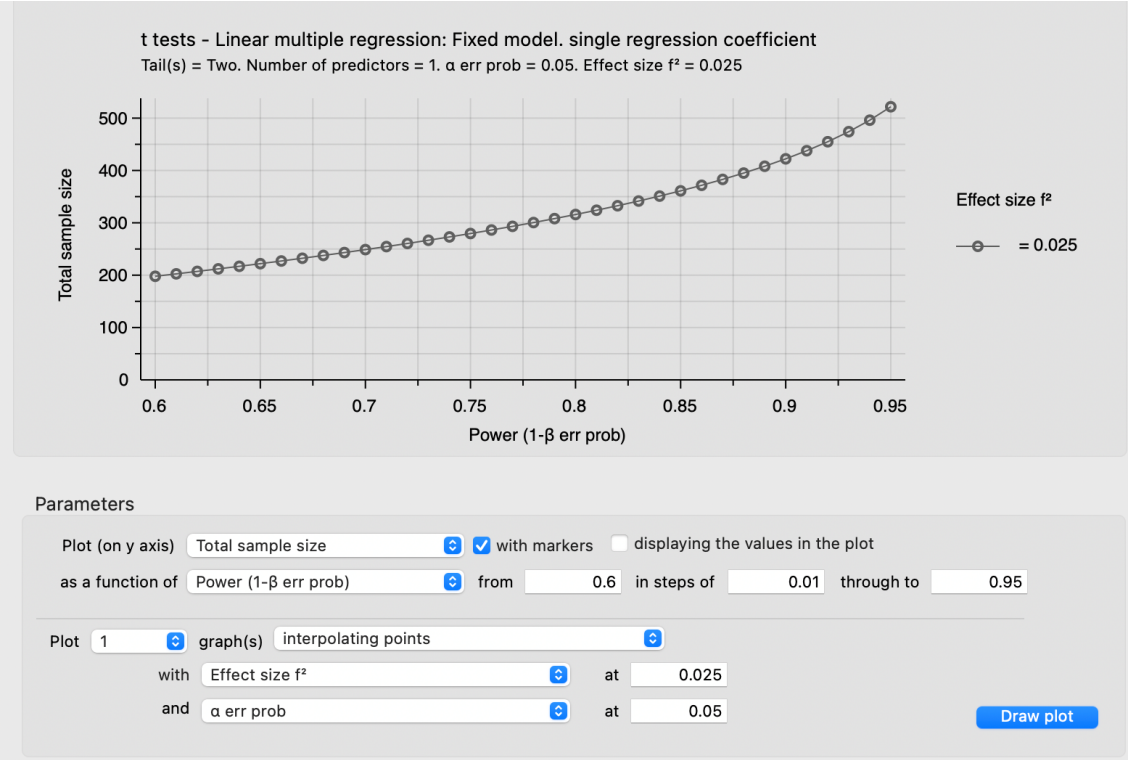
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials from 2018–2021. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A5: Test Statistics Distribution by Pre-Registration Status: Derounding



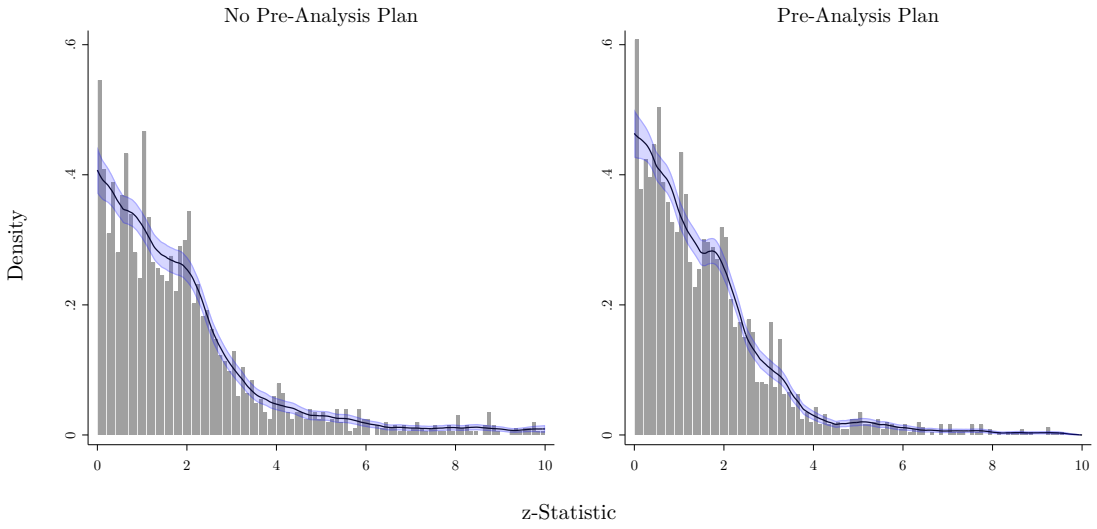
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials from 2018–2021 by pre-registration status. We define a pre-registered RCT as a study that was registered before its trial end date listed in a registry. Studies that were registered after the trial end date are counted as non-pre-registered. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We use the inverse of the number of tests presented in the same article to weight observations.

Figure A6: Power Test



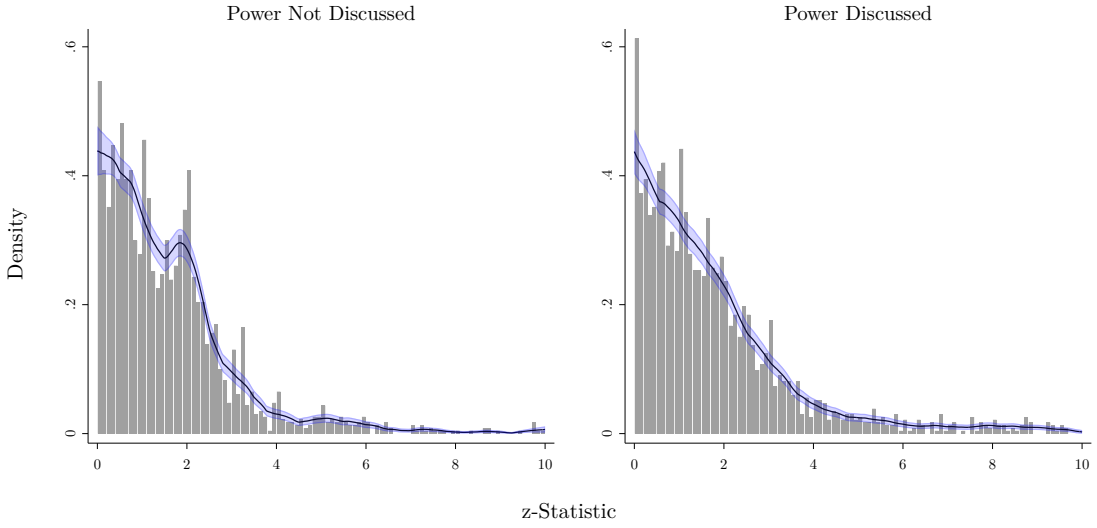
Notes: This figure shows statistical power to detect an effect of 0.025. We expect that our sample size for the caliper test using $z \in [1.46, 2.46]$ is over 6,000. We also expect that approximately 33% of test statistics are in pre-registered RCTs.

Figure A7: Test Statistics Distribution for Pre-Registered RCTs by a Presence of Pre-Analysis Plan: Derounding



Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials that were pre-registered from 2018–2021 by presence of a pre-analysis plan. We define a pre-registered RCT as a study that was registered before its trial end date listed in a registry. Studies that were registered after the trial end date are counted as non-pre-registered. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

Figure A8: Test Statistics Distribution for Pre-Registered RCTs by Presence of Power Analysis: Derounding



Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials that were pre-registered from 2018–2021 by power analysis status. We define a pre-registered RCT as a study that was registered before its trial end date listed in a registry. Studies that were registered after the trial end date are counted as non-pre-registered. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

12 Appendix Tables

Table A1: Summary Statistics by Journal

Journals	Articles	Tests	Prop. Articles Pre-Registered
	(1)	(2)	(3)
American Economic Journal: Applied Econ.	34	2,284	0.19
American Economic Journal: Econ. Policy	10	447	0.06
American Economic Review	36	2,106	0.63
Econometrica	4	97	0.00
Economic Journal	14	891	0.18
Journal of Development Economics	77	4,318	0.20
Journal of Finance	2	67	0.00
Journal of Human Resources	17	847	0.30
Journal of Labor Economics	6	185	0.05
Journal of Political Economy	14	949	0.63
Journal of Public Economics	42	1,501	0.17
Journal of the European Econ. Association	10	264	0.31
Quarterly Journal of Economics	22	1,287	0.53
Review of Economic Studies	12	513	0.02
Review of Economics and Statistics	14	584	0.35

Notes: This table alphabetically presents our sample of journals. We report the number of articles and tests per journal in this table. The last column show the percentage of studies that were pre-registered. Note that the numbers in column 1 are based on the number of articles for the years 2018–2021.

Table A2: Caliper Test, Statistically Significant at the 5 Percent Level: Logit

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	-0.018	-0.026	-0.024	-0.013	-0.021
	(0.027)	(0.026)	(0.027)	(0.029)	(0.035)
Controls					
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Articles' Charact.		Y	Y	Y	Y
AEA & Top 5		Y			
Journal FE			Y	Y	Y
Observations	3,870	3,779	3,779	2,789	1,671
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]

Notes: This table reports marginal effects from logit regressions. The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table A3: Caliper Test, Statistically Significant at the 10 Percent Level

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	-0.018	-0.026	-0.023	-0.013	-0.021
	(0.027)	(0.026)	(0.027)	(0.029)	(0.035)
Controls					
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Articles' Charact.		Y	Y	Y	Y
AEA & Top 5		Y			
Journal FE			Y	Y	Y
Observations	3,870	3,779	3,779	2,789	1,671
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]

Notes: This table reports marginal effects from probit regressions. The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table A4: Caliper Test, Statistically Significant at the 1 Percent Level

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	0.003	0.006	-0.004	-0.042	-0.028
	(0.032)	(0.036)	(0.033)	(0.038)	(0.049)
Controls					
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Articles' Charact.		Y	Y	Y	Y
AEA & Top 5		Y			
Journal FE			Y	Y	Y
Observations	2,410	2,368	2,368	1,572	899
Window	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]	[2.58±0.35]	[2.58±0.20]

Notes: This table reports marginal effects from probit regressions. The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table A5: Caliper Test, Statistically Significant at the 5 Percent Level: Article Weights

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	-0.029	-0.041	-0.041	-0.018	-0.026
	(0.029)	(0.026)	(0.028)	(0.031)	(0.039)
Controls					
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Articles' Charact.		Y	Y	Y	Y
AEA & Top 5		Y			
Journal FE			Y	Y	Y
Observations	3,870	3,779	3,779	2,789	1,671
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]

Notes: This table reports marginal effects from logit regressions. The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.