INSTITUTE for REPLICATION

No. 101
I4R DISCUSSION PAPER SERIES

Do Pre-Registration and Pre-Analysis Plans Reduce p-Hacking and Publication Bias?
Evidence from 15,992 Test Statistics and Suggestions for Improvement

Abel Brodeur

Nikolai M. Cook

Jonathan S. Hartley

Anthony Heyes

January 2024

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 101

Do Pre-Registration and Pre-Analysis Plans Reduce p-Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement

Abel Brodeur[1], Nikolai M. Cook[2], Jonathan S. Hartley[3], Anthony Heyes[4]

[1]University of Ottawa/Canada
[2]Wilfrid Laurier University, Waterloo/Canada
[3]Stanford University, Stanford/USA
[4]University of Birmingham/Great Britain

JANUARY 2024

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Do Pre-Registration and Pre-Analysis Plans Reduce p-Hacking and Publication Bias?: Evidence from 15,992 Test Statistics and Suggestions for Improvement

Abel Brodeur Nikolai M. Cook Jonathan S. Hartley
Anthony Heyes
University of Ottawa, Wilfrid Laurier University, Stanford University
University of Birmingham

January 12, 2024

Abstract

Pre-registration is regarded as an important contributor to research credibility. We investigate this by analyzing the pattern of test statistics from the universe of randomized controlled trials (RCT) studies published in 15 leading economics journals. We draw two conclusions: (a) Pre-registration frequently does not involve a pre-analysis plan (PAP), or sufficient detail to constrain meaningfully the actions and decisions of researchers after data is collected. Consistent with this, we find no evidence that pre-registration in itself reduces p-hacking and publication bias. (b) When pre-registration is accompanied by a PAP we find evidence consistent with both reduced p-hacking and publication bias.

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

JEL CODES: B41, C13, C40, C93.

Authors: Brodeur: University of Ottawa. 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 are grateful to Isaiah Andrews, Abhijit Banerjee, Nina Buchmann, Arun Chandrasekhar, Pascaline Dupas, Marcel Fafchamps, Peter Hull, Guido Imbens, John List, Steve Levitt, Eva Lestant, Neil Malhotra, Melanie Morten, Muriel Niederle, Ben Olken and three referees from this journal for very helpful advice. Abigail Marsh and Susan Price provided excellent research assistance. The pre-analysis plan (https://osf.io/q9bxm/) for our study 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. Our study should be regarded as exploratory since we cannot evidence definitively that the PAP was posted prior to conducting any analyses. Heyes acknowledges financial support from the Canada Research Chair programme. Errors are ours. Edited by Anna Dreber.

1 Introduction

RCTs have become increasingly prominent in economics, and the social sciences more broadly, in recent years. By combining deliberate, researcher-controlled 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 increase has coincided with a wider trend towards open science practices and research transparency in the profession, aimed at bolstering the credibility of the results of empirical research. In the particular case of RCTs, central to this has been the promotion of the 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.

Registration on the AEA RCT Registry is required for all RCTs submitted to AEA journals, and J-PAL is among the funders that encourage pre-registration for trials. However, the content of such pre-registration can vary substantially. Though the option exists for researchers to provide a high level of detail on how they plan to proceed, the elements that are *required* by the platform are skeletal. The final field in the submission window invites the inclusion of a pre-analysis plan (PAP). A PAP is a more detailed and explicit statement of which hypotheses are to be tested and how. Proponents would claim that inclusion of a PAP can reduce meaningfully a researcher's 'wiggle' room after data has been collected, reducing the extent of p-hacking by limiting the scope for a researcher to change her analytical choices after statistical significance has been observed. A PAP may also reduce publication bias by making selective reporting of results more difficult.

Importantly, however, the applicant for pre-registration at the AEA Registry decides whether or not to include such a plan (about half do) and, as such, pre-registration and provision of a PAP are distinct and separable things. This approach

to the meaning of pre-registration is quite different to those in other disciplines, notably psychology, where pre-registration implies a PAP by definition. This separability likely contributes to ongoing confusion about what pre-registration, as practiced in economics, implies. A further layer of ambiguity is provided by recognizing that (a) PAP's uploaded to the AEA Registry are neither vetted nor required to meet minimum standards, so can vary in stringency, while (b) the rest of the portal includes several non-PAP fields in which authors could, if they wished, provide a detailed description of their methods, data collection, definition of dependent and independent variables, etc. in a way that, if populated comprehensively, could amount to a PAP.

It seems likely that much of this ambiguity in what pre-registration *as economists practice it* implies is not well-understood by many in the profession. The uninformed may only note that the study contains the recommended standard-issue statement of status, "(T)his study is registered in the AEA RCT Registry under ID number doi.org/10.1257/rct.1234", which makes no reference to the inclusion or otherwise of a PAP, the precision or vagueness of the pre-registration information, including any PAP that does exist, and the variation between what is pre-registered and what features in the published manuscript.

In this paper, we investigate *separately* the relationship between pre-registration, the vagueness of pre-registration, the existence of a PAP and patterns of statistical significance. In particular we look for evidence of associations with reduced p-hacking (i.e., manipulation and/or selective reporting of results' p-values) and/or publication bias (i.e., the statistical significance of results in a study influencing the probability that a result is published). Both p-hacking and publication bias make the body of published evidence less credible or trust-worthy.[1] It is worth observing that it is yet to be established what the correct relationship should be between what is contained in a PAP, if one is provided, and the published paper. Interestingly, Banerjee et al. (2020) proposes that researchers could consider producing a short,

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

publicly available report that populates the PAP to the extent possible, in addition to a fully-fledged paper that may digress from what was originally planned. The standard in many general science and psychology journals that authors make clear any departures from pre-registration in the main text has not evolved as a norm, nor been imposed by journals or professional associations, in the discipline of economics.

It is worth underlining here that we proceed with notions of pre-registration and PAPs *as practiced in economics*, or at least as operationalized by the largest and most influential professional association in the discipline, the AEA. In their discussion relating to psychology, Nosek et al. (2018) contend that: "An effective solution is to define the research questions and analysis plan before observing the research outcomes – a process called preregistration," which implies that pre-registration and the existence of a PAP are one and the same thing (see also Simmons et al. (2021) for a similar contention). This is far from how things work in economics, as we will show here. We contend that many readers believe, wrongly according to our analysis, that pre-registration in itself implies enhanced research credibility, which would explain the weight apparently attached to whether a study is pre-registered or not in assessing its likely validity. Our finding is that credibility is enhanced only with inclusion of a PAP.

To do this we harvest and analyze the universe of hypothesis tests drawn from RCTs published in 15 leading economics journals in the years 2018 through 2021. The analysis includes 314 journal articles, of which 83 are treated as pre-registered, and 15,992 test statistics. Of the 83 pre-registered articles, 44 have a PAP (39 are pre-registered but do not have a PAP).

As a first step, for descriptive interest, we show that the use of both pre-registration and PAPs 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 were pre-registered in 2018 compared to 40% in 2021 (Figure 1).

We next investigate the relationship between article and author 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 no evidence that more experienced or elite scholars are any more/less prone to adopt and use pre-registration. We do, however, find that rates of pre-registration vary widely between journals, with the very top journals having particularly high pre-registration rates.

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 things 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 on either side of conventional significance thresholds. Second, 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. Last, we check the robustness of our results by applying popular tests for the detection and measurement of *p*-hacking and publication bias devised by Brodeur et al. (2020) and Elliott et al. (2022). There is no definitive way to discern *p*-hacking or publication bias and we believe that our approach of using multiple methods, each requiring it's own assumptions and subject to its own limitations, allows for a more rounded assessment.

Across the analyses we find no evidence that pre-registration *in itself* is associated with reduced *p*-hacking or publication bias when focusing on the 5% and 10% significance thresholds. This may surprise some readers many of whom - anecdotally at least - may attach substantial weight to whether a particular study is pre-registered or not in making an evaluation.

Next we explore the additional role of the inclusion of a PAP in the pre-registration. We find that inclusion of a PAP is associated with less *p*-hacking and publication bias, at least in the vicinity of the 5% significance threshold, which we believe to be that most salient to economists. Furthermore, pre-registered RCTs that provide an explicit discussion of statistical power/sample size are less prone to

p-hacking than those that do not provide such a discussion. We then investigate in more detail the role of pre-registration vagueness.

We conclude that the extent of p-hacking in well-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.[2] In terms of possible policy implications - for example publishers or funders requiring the use of PAPs - we acknowledge a variant on Goodhart's Law, that while looking backwards increased credibility was correlated with use of PAP, such a correlation would not necessarily sustain in a setting in which PAPs were mandated.

Our study adds flesh 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. (Forthcoming); Butera et al. (2020); Camerer et al. (2019); Casey et al. (2012); Christensen et al. (2019); Drazen et al. (2021)).

Abrams et al. (2020) observe that most RCTs in economics (90 percent in their sample) are not registered, and based upon an audit of pre-registrations contend that most are not sufficiently detailed to significantly aid inference. Our results complement theirs since, in contrast to our study, they do not examine the pattern of statistical significance in the resulting body of research results. Readers in economics appear to attach substantial weight to a study being pre-registered, even though, as Abrams et al. (2020) establish, pre-registration in itself places little restriction on subsequent practice.

Four additional studies are especially pertinent and worth noting here. 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, find-

[2]Given that the processes of pre-registration and pre-analysis are costly whether the benefits in improved credibility justify the costs researcher involved are not questions that we address here.

ing 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, though of course they could be required to be more comprehensive. Oostrom (2022) shows that pre-registration requirements for drug efficiency trials mitigate sponsorship bias. 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 among the first to document the relationship between p-hacking, publication bias and pre-registration in economics. Other meta-analyses in the social sciences include Scheel et al. (2021) and Lewis et al. (2022) and Kvarven et al. (2020). In their study of effect sizes, Kvarven et al. (2020) compare the results of meta-analyses to large-scale, multi-laboratory replications of 15 results in psychology. This is an attractive setting since replications provide precisely estimated effect sizes that do not suffer from publication bias or selective reporting. They find that effect sizes from meta-analyses are significantly larger than those from replications in 12 of the 15 cases, on average by a factor of three.

Our study also contributes to the 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); Brodeur et al. (2023); Doucouliagos and Stanley (2013); Furukawa (2019); Gerber and Malhotra (2008); Havránek (2015); Havránek and Sokolova (2020)).[3] 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

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

of causal inference. In psychology, Scheel et al. (2021) provide evidence that registered reports - a form of journal article in which peer review of the study protocol and the decision to publish occur before the study is run - increase the reporting of "negative" (not supportive of tested hypothesis) results.

Overall our findings point to several ways in which economists could consider changing practices around pre-registration, each of which would bring them closer to standards in other disciplines where experimental methods are frequently used, notably psychology. First, making a PAP a compulsory element of pre-registration, and ensuring that the PAP meets certain standards of detail. Second, encouraging or requiring that authors make explicit in their manuscript text any deviation from the PAP. The latter ensures that results from exploratory strands of research, developed after data is collected, can still be shared, but readers are informed to their non-pre-committed character. We will return to draw out some possible implications of the analysis in more detail in the conclusions.

2 Conceptual Framework

Before getting to the data we sketch briefly why PAPs and pre-registration might be expected to decrease, to differing degrees, p-hacking and/or publication bias. More complete discussions of (some of) the advantages and disadvantages of pre-registration and PAP usage in economics have been provided by Banerjee et al. (2020), Coffman and Niederle (2015), and Olken (2015), among others.

It is apparent that any mechanism that obliges a researcher to 'commit' to the hypotheses he or she will test, and how he or she will process the data in order to test those hypotheses, can potentially reduce the prevalence of p-hacking. The extent of the discipline imposed will depend upon the level of detail incorporated, and the rigor with which adherence is enforced. As already noted, basic pre-registration of a study requires only rather coarse information on intent, whereas a complete PAP would include the econometric specifications, outcome variables, cleaning procedures and other methodological details that could ultimately influence statisti-

cal significance. If the regressions are pre-specified and researchers report (or are required to report) all the results pre-specified, p-hacking would be expected to become much less of a problem.[4]

A related benefit of pre-registering a complete PAP is a potential reduction in the extent of publication bias, at least at the working paper stage. Disciplined by a PAP, researchers are plausibly more likely to report null or negative results, following their commitment in advance to so doing. Pre-registration may also prevent or make harder to formulate or add hypotheses 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 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] Those costs would be further elevated if the standards for the level of detail that is required in a PAP for it to be deemed adequate were enhanced.

In addition, in some cases it might also not be possible strictly to follow a PAP,

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

[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.

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, new outcomes of interest might arise, or new statistical techniques become available. It might very well be undesirable to prohibit publication of results from any such analyses, but any variation between the original pre-registration and what is included in the published version could be clearly identified to readers.

3 Data

We focus on leading economics 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). Note that our full sample includes 314 titles, of which 83 are treated as pre-registered. Of these 83 articles, 44 have a PAP (39 are pre-registered but do not have a PAP).[8] Noting that authors and journals vary with respect to the precision with which estimates are reported, we collect all decimal places.

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

[8] In our pre-analysis plan, we stated that "Our final sample should include over 400 journal articles. We anticipate that our final sample will include about 20,000 test statistics."

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 graduated; and whether they are editors of an academic journal at the time of publication. The latter information is collected from author websites and curriculum vitae available online. We code top institutions using the highest rated 20 in RePec's ranking of top institutions.[9]

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*, preregist*, pre-analysis plan, socialscienceregistry`, and `PAP`. *aearct* stands for the AEA RCT Registry and *osf* stands for the Open Science Framework. 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-

[9]The following 20 institutions are 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.

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.[10]

We present summary statistics in Tables 1 and 2. In Table 1, we report the mean 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 any academic journal at the time of publication. About 30% of articles are published in 'Top 5' journals.[11] Appendix Table A1 provides a breakdown by journal.

Around 30% of the test statistics in our sample come 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* (Ap-

[10]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 has no impact on our conclusions in a set of robustness exercises.

[11]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*.

pendix 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 (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.[12] 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 ($p = 0.017$ in column 6). 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 ($p < 0.000$ and $p = 0.003$).

Experience does not seem to play an important role in determining pre-registration ($p = 0.961$) nor does current affiliation ranking ($p = 0.777$). We do not find evidence that authors graduating from top institutions are more likely to pre-register, as the estimates are statistically insignificant at the 5% level ($p = 0.441$). 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 ($p < 0.000$). Of note, this relationship should not be viewed as causal. 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

[12] See Appendix Table A2 for linear probability models.

measure p-hacking, including caliper tests which allow us to control for confounders.

5 Pre-Registration, *p*-Hacking and Publication Bias

Here we describe our graphical and formal analyses to detect and measure p-hacking and publication bias. We rely on two methods comparing the distribution of test statistics by pre-registration status. We also provide additional tests in the appendix. 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.

We first wish to note a limitation of our analysis: none of the methods used enable us cleanly to identify p-hacking. Visual inspection, calipers, and the battery of tests proposed in Elliott et al. (2022) all jointly identify p-hacking and publication bias in the presence of either (or both). A further limitation is that p-hacking and publication bias have been shown to lead to inflated effect sizes (Gelman and Carlin (2014); Ioannidis (2008)), which our methods do not detect nor quantify. A third limitation is a focus on marginal p-hacking and publication bias (where statistical significance is just or just-not achieved). Non-marginal p-hacking and publication bias, well outside the unit interval surrounding our statistical significance thresholds could be occurring without detection by the applied methods here. Nonetheless, we apply what we believe to be the state of the art, and remain hopeful that future research develops a clean identification of non-marginal p-hacking and publication bias.

5.1 Reporting

One 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 might be that it makes it harder for hypotheses

to be discarded once results are seen.

We find some evidence consistent with that in our sample. Pre-registration is associated with researchers reporting 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 ($p = 0.050$). 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.

5.2 Graphical Analysis

Here we 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. The discernible 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 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 the other causal identification methods of empirical work used in these economics journals (Brodeur et al. (2020); Vivalt (2019)).

We now turn to the question of whether the distribution of test statistics varies by pre-registration status. This graphical analysis will provide us a first (informal) test of whether pre-registered studies are less p-hacked. 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.[13] 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 ($p = 0.125$).[14]

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 now 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. A fourth explanation is the presence of omitted variables driving such a relationship. We come back to this issue in the next subsection.

Overall, our graphical analysis offers little evidence that pre-registration reduces the forms of p-hacking considered in our analysis.

[13] 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.

[14] 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.

5.3 Caliper Tests

Caliper tests are well-established tools in research on p-hacking and publication bias, and involve comparing the number of test statistics in a narrow equal-sized range above and below arbitrary significance thresholds. If there is a large difference in the number of observations just above a statistical significance threshold, we take this as evidence towards the presence of p-hacking or publication bias. Of note, the caliper test method jointly identify p-hacking and publication bias in the additional presence of publication bias.

The main advantage of this method is that we can control for confounders in our estimation. This is key in our setting as authors who pre-registered their RCTs have different characteristics. 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 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.[15] 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 Tables 4 and 5 for the 5% and 10% statistical significance thresholds, respectively. (See Appendix Table A9 for the 1% significance threshold.) In columns 1–3, we restrict the sample to $z \in [1.46, 2.46]$ for

[15] Our results are robust to using logit models, OLS and derounding. See Appendix Tables A3, A4, A5, A6, A7 and A8.

the 5% statistical significance threshold. Our sample size for the caliper test using $z \in [1.46, 2.46]$ is 3,870.[16] 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 RCTs that were not pre-registered. The point estimate is very small (-0.020) and statistically insignificant at conventional levels ($p = 0.449$).

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 (lowest $p = 0.262$). The results are similar for the other statistical significance thresholds.[17]

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

We now 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 one of Andrews and Kasy (2019)'s methods for identifying the conditional probability of publication as a function of a study's results. Their first method involves systematic replication studies and the second on meta-studies. We use the second method. The primary estimated parameter of interest to our study provided by Andrews and Kasy (2019) is the relative publication probability of a statistically

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

[17] Our results are also robust to weighting. In Appendix Table A10, we use the inverse of the number of tests presented in the same article to weight observations.

significant result compared to a statistically *in*significant result. The results are reported in Table 6. 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 only 0.60 times as likely to be published as a statistically significant one.) This is important degree of publication bias, although modest compared to that in some other parts of the empirical economics literature.[18] The estimated parameter for pre-registered RCTs is 1.65, for non-pre-registered RCTs 1.67. In other words, perhaps surprisingly, pre-registration status seems to have no meaningful impact on publication bias.

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

5.5 Other Tests for Detecting and Measuring the Extent of *p*-Hacking Proposed by Brodeur et al. (2020) and Elliott et al. (2022)

We now show that our results are robust to the use of other tests to detect and quantify the extent of p-hacking. We provide more details in the Online Appendix.

First, we rely on 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 that we would expect to emerge absent publication bias. We provide much more details in Appendix 12.1 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%

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

statistical significance threshold where there are 10% more than expected.

Second, 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 eight tests against a null hypothesis of no p-hacking. See Appendix 12.1 for more details. 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 three 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.

Overall, these additional tests confirm our previous results suggesting that pre-registration in itself does not appear to play an important role in reducing p-hacking.

6 Pre-Analysis Plans (PAPs) and Pre-registration Vagueness

So pre-registration in itself seems to have no meaningful impact on the tendency to p-hacking and publication bias in the RCT literature.

Here we turn to the important question of whether (a) the presence of a PAP and (b), the degree of vagueness in pre-registration, makes a difference. (The analysis in this section was not pre-registered and should be viewed as exploratory. We followed suggestions from seminar participants and other researchers listed in the acknowledgement footnote.)

6.1 Presence of a PAP

To explore the additional discipline imposed by pre-registration of a PAP we exploit that the AEA RCT repository provides the opportunity for authors to pre-register a fully-fledged "Analysis Plan", without requiring it. As such, 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", etc.).[19] We explore some of these dimensions in what follows.

For each paper we code the presence of a 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 coded as having a PAP. If the registration was made after the end trial date, it would not be counted as having a PAP nor a pre-registration.

Before illustrating the distribution of test statistics for our different subsamples, we first relate use of PAP to observable author and article characteristics. For this we rely on probit regressions and estimate equation 1 where the dependent variable is a dummy variable for whether test i in journal article a in journal j has a PAP. We restrict the sample to pre-registered RCTs. The findings are presented in Appendix Table A11.[20]

Overall, most of the point estimates are statistically insignificant at conventional levels. We find evidence that authors graduating from top institutions are less likely to provide a PAP ($p = 0.004$).

We now test whether the extent of publication bias and p-hacking differ for subsamples of pre-registered RCTs with and without a PAP. Figure 4 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.[21] Both curves are monotonically falling with a bump around the 5% significance threshold. Notably,

[19] A pre-registered PAP could lead to a greater number of tests being reported. For example, results for main hypotheses are likely to be reported by both studies with and without a PAP. However statistically insignificant tests of secondary or sub-results must be reported if they feature in a PAP, whereas in the absence of a PAP the researcher may opt to discard them. (see Olken (2015) for a discussion). We do not find any evidence of this effect; we collected 62 test statistics per article for pre-registered studies with PAPs in comparison to 56 for those without a PAP. The difference is not statistically significant, though we note that the number of pre-registered articles with PAPs is relatively small.

[20] See Appendix Table A12 for the same analysis but for the full sample (i.e., removing the restriction of pre-registration).

[21] See Appendix Figure A7 for the corresponding derounded distributions.

visually the bump is more pronounced for the curve without a PAP and the curve for PAPs contains more tests with large p-values.

Our caliper tests confirm these patterns. In Tables 7 and 8, we replicate the structure and specifications of Tables 4 and 5, replacing the dummy variable pre-registration by a dummy for whether the test statistic is in an RCT containing a PAP.[22] We restrict the sample to pre-registered RCTs.[23] We have 1,164 and 1,278 observations for our baseline window at the 5% and 10% significance levels. For the 5% significance threshold, our estimates are all negative and suggest 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 out of five estimates are statistically significant at the 5% level, including all our specifications with journal fixed effects ($p = 0.012$ in column 6). In contrast, we find no difference in the proportion of test statistics that are marginally significant in articles with and without a complete PAP at the 10% significance threshold.

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 6). 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.

Last, we find little evidence that having a pre-registered PAP is associated with researchers reporting more results. We collected 62 tests per article for pre-registered studies with PAPs in comparison to 56 for those without a PAP. The difference is not statistically significant at conventional levels ($p = 0.908$).

[22] Appendix Table A13 replicates Table 7 but omits additional variables discussed below for the completeness of the pre-registration. The point estimates are strikingly similar. The results are also robust to derounding. See Appendix Table A14.

[23] We replicate these tables including our full sample in Appendix Tables A15 and A16. These tables confirm our result that pre-registered RCTs with a PAP are significantly less likely to reject the null in comparison to those without a PAP, but not in comparison to non-registered RCTs. This could be due to selection into pre-registration and omitted variables.

6.2 Pre-Registration Vagueness

We now investigate whether the 'vagueness' of pre-registration is related to *p*-hacking and publication bias.

We first detail our methodology. We manually read the content of registrations for pre-registered RCTs in our sample. For each RCT, two authors manually read and compared the "primary outcomes" variable as defined in the pre-registration to outcomes reported in the published article. A total of 4,893 tests are associated with a pre-registration, approximately 30% of our sample.[24]

We classify coefficients of interest in each article (see Section 3 for details) into a binary variable; whether the estimate represents a test that was pre-registered or not. Said differently, for each test, we code whether the dependent variable was included in the pre-registration. Each classification was conducted by at least two of the authors to minimize coding errors.[25]

We note here a potential complication and one remaining at the core of the issue with pre-registration as it is currently practised in economics. An author, aware they must pre-register their research in order to publish it in many top-outlets, may do so in a number of ways. The (perhaps unrealistic) ideal would be a small number of outcome variables with sufficient detail so that their measurement and definitions could be reproduced by an independent researcher.

We find that a large number of pre-registrations do not conform to such an ideal. Many pre-registrations are written in a 'vague' manner. For example, the primary outcome may be defined as "health outcomes" as compared to "incidence of cardiac arrest". While remaining agnostic about the motivations of the researcher who completes a vague pre-registration, but note that the potential for p-hacking after results are known is likely much lower for those with a specific pre-registration. Sep-

[24]The AEARCT allows for publication embargoes, which prevent accessing the details of the pre-registration until after a certain date; there were no instances where we were unable to access the details of a pre-registration in our sample (presumably because authors are correctly releasing their embargoes following publication).

[25]There were a small number of initial disagreements in coding whether the dependent variable was included. In such cases, we would discuss among ourselves and make a decision. Excluding these tests has no impact on our conclusions.

arately, we find a significant number of articles that contain pre-registered outcome variables that are not referred to at all in the final published article. While these pre-registrations are not entirely vague (and it is commonplace for many RCTs to collect a large number of outcomes for legitimate research purposes) it is possible for the researchers to circumvent the credibility benefits of pre-registration by including a broad pre-registration.[26] Therefore, we coded a second variable that corresponded to the vagueness of outcomes contained in the pre-registration. This variable was also coded by two authors and any disagreements discussed at length.[27]

We find the following results. About one-third of estimates in pre-registered articles correspond to outcome variables not described in the pre-registration. A total of eight articles contain no pre-registered coefficients of interest, while 40 contain only pre-registered coefficients of interest. The remaining 35 articles had both pre-registered and non-pre-registered tests.

Unsurprisingly, the share of estimates with an outcome pre-registered is higher in vague pre-registrations (74%) than in a non-vague (henceforth specific) pre-registrations (68%).[28]

Figure 5 illustrates the distribution of test statistics for pre-registered RCTs for (i) tests using a dependent variable that was pre-registered with a specific description of the outcomes; (ii) tests using a dependent variable that was pre-registered with a vague description of the outcomes; (iii) tests not using a dependent variable that was pre-registered with a specific description of the outcomes; and (iv) tests using a dependent variable that was not pre-registered with a vague description of the outcomes, respectively.[29] All curves are monotonically falling with a small

[26]While the AEARCT registry offers a dedicated field for 'secondary outcomes', we do not distinguish between primary and secondary outcomes in our analysis. This field for secondary outcome was used only four times in our sample of 83 pre-registrations.

[27]See https://osf.io/v6eqc for an example of a comprehensive and specific description of the outcome variables in the pre-registration. A brief but precise description of the outcome variables would also be coded as specific (e.g., outcome variables are test scores in math and language). In contrast, listing all or some of the outcome variables loosely such as "female empowerment" or "business practices" was coded as vague.

[28]A simple regression of table number against pre-registration confirms that pre-registered results are most often presented earlier in an article (more than one half or 0.61 tables earlier than average, on a mean 3.8 tables and standard deviation of 1.8).

[29]See Appendix Figure A8 for the corresponding derounded distributions.

bump around the 5% significance threshold, with the exception of subfigure (iv) which is not downward-sloping over the support $z < 2$ and has relatively more mass around 1.96.

We test formally whether this result holds when using caliper tests in Tables 7 and 8.[30] We include the following variables; (i) a dummy variable for whether the test uses a dependent variable that was pre-registered, (ii) a dummy variable for whether the pre-registration is specific, and (iii) an interaction term. We find no evidence that the specificity of the pre-registration is related to the extent of p-hacking ($p = 0.268$). We only provide suggestive evidence (for the 10% significance threshold) that having a specific pre-registration and pre-registered outcomes is related to a smaller extent of p-hacking ($p = 0.050$). The estimates for the interaction term are always positive but statistically significant only in our full model using the smallest window.

Turning to Andrews and Kasy (2019)'s method for the estimation of publication bias, we find that the extent of publication bias is higher for tests using a dependent variable that was not pre-registered with a vague description of the outcomes in comparison to the other three categories (see Table 6). Other things equal, a statistically significant result is estimated to be 1.90 times *more* likely to be published than a statistically insignificant result for this group. This compares with 1.71 times for tests using a dependent variable that was pre-registered with a specific description of the outcomes.

6.2.1 Discussion of Statistical Power Our focus throughout this study has been on patterns of statistical significance, and sample size is central to that. An important aspect of pre-registration precision, therefore, is the explicit treatment of sample size.

To investigate this we create a variable, following Ofosu and Posner (2021), for whether the PAP included a power analysis.[31] Figure 6 shows the distribution of

[30] See Appendix Table A17 for OLS estimates.

[31] 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

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.[32]

Visually, we find that the distributions are very different (a two-sided Kolmogorov–Smirnov test also rejects the null of equality of distributions with $p < 0.000$). There is a substantial bump around the 5% significance threshold for articles that do not include an explicit discussion of statistical power and a monotonically falling curve with no such bump, for RCTs including such a discussion.

Andrews and Kasy (2019)'s methodology suggests that a statistically significant result is 1.77 times *more* likely to be published than a statistically insignificant result for articles that do not include an explicit discussion of statistical power in comparison to only 1.38 times for those that include such a discussion. In contrast, the caliper tests do not reveal that the inclusion of a discussion for statistical power is related to p-hacking (see Table 7).[33]

To sum up, we provide some evidence that discussion of statistical power may reduce the extent of publication bias, but not p-hacking.

7 False Discovery Rates

Finally, we turn briefly to the role of false discovery rates. In particular, we investigate whether authors that report q-values for false discovery rates are less likely to suffer from p-hacking and publication bias. Using and reporting q-values may be an interesting tool for RCTs involving a large number of dependent variables. We flag articles where the text of the article contains one or more of the following

the authors specify a clear hypothesis; specify the primary dependent and independent variables sufficiently clearly to preclude post-hoc adjustments; spell 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.

[32] See Appendix Figure A9 for the corresponding derounded distributions.

[33] Ioannidis et al. (2017) rely on meta-analyses for which a comparatively small number of point estimates is collected per study. In contrast, we collect all point estimates of interest for each study. Our study cannot use the method proposed by Ioannidis et al. (2017) since in our study we are mixing many different strands of research, relating to many different prospective effects. In contrast, Ioannidis et al. (2017) has the main point estimate for each meta-analysis and can thus estimate power for each meta-analysis. In addition, our sample deals with RCTs, while most of the sample in Ioannidis et al. (2017) is non-experimental.

keywords: `q value, false discovery`, and `fdr`. We then read the articles and confirm that the authors did report q-values. We identify 18 RCTs using q-values in our sample.

Figure 7 shows the distribution of test statistics for articles reporting q-values (right panel) and those without (left panel), respectively.[34] Visually, again, we find that the distributions are very different, with a discernible bump around the 5% significance threshold for articles that do not include q-values and a monotonically falling curve without a bump for RCTs relying on false discovery rates. This result is confirmed with Andrews and Kasy (2019)'s method, while we find little evidence using caliper tests (Appendix Tables A18 and A19). The estimates are statistically insignificant in most columns, with the magnitude and sign changing once we add control variables. It is not possible to say for sure why this is the case, but it suggests that RCTs relying on q-values may be different in scale or nature relative to those that do not.

8 Conclusion

The credibility of the published body of empirical research in the social sciences has come under increasing scrutiny 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 to mitigate such credibility concerns, with particular emphasis put on pre-registration of studies. The aim of the current study is 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.

Studying the universe of RCTs published in 15 leading economics journals over a four year period, we find that pre-registration *in itself* is not a correlate of increased credibility - there is no discernible difference between the pattern of test

[34] See Appendix Figure A10 for the derounded distributions.

statistics found in pre-registered studies and those from their non-pre-registered counterparts when focusing on either significance threshold. However, the inclusion of a PAP *is* associated with less p-hacking and publication bias. These results point to a potentially important mitigative effect of pre-registration when combined with PAPs.[35]

It is important to repeat that the analysis here relates to pre-registration *as practiced in economics*. As we have already noted, pre-registration in other fields, notably psychology, is quite different in character, typically requiring the inclusion of a PAP. Given the apparent importance attached in the profession to a study being pre-registered, it seems likely that many do not recognize how skeletal the minimum requirements for pre-registration in, for example, the AEA Registry are, and how little restriction most pre-registrations places on researcher behavior.

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 one plausibly not made randomly. While we can document the characteristics of researchers correlated with this research practice drawing causal conclusions is more challenging. An interesting question for future research would be whether or to what extent researchers use pre-registration strategically, for example proceeding with what they know to be a minimalist pre-registration to benefit from the additional validity that the pre-registration credential appears to deliver, without any additional restriction on behavior.

Being careful not to over-interpret results, drawn as they are from a narrow subset of journals, with some parts reliant on relatively small sub-samples, our broad conclusion is that the process of pre-registration *as currently practiced in economics*, for example as operationalized by the AEA Registry, does not enhance the reliability of research results. This may appear a straw-man result, since the most skeletal pre-registration that allows the researcher to assert that '(T)his study is registered in the AEA RCT Registry under ID number doi.org/10.1257/rct.1234'

[35] Another potential benefit of pre-registration is to lower inflation of effect sizes (Strømland (2019)), which have been documented in a number of disciplines (e.g., Ioannidis (2008)).

places no meaningful restriction on research conduct. However, we believe the minimal requirements for pre-registration are not well understood by most in the profession. This provides the possibility that its operation is counter-productive, with 'cheap talk' pre-registration to have an important impact on how research is perceived. An ambition of future research is to evidence that claim.

This is not to say that pre-registration *done differently* could not enhance research credibility. We show that the inclusion of a PAP in the pre-registration predicts such enhancement, albeit in a limited sample, and that non-vagueness in certain elements of specification is also a correlate.

To that end several recommendations for reform of pre-registration practices can be stated.

First, if pre-registration is to become a standard in the profession it should be operationalized in a way that places meaningful restrictions on the discretion researchers have to make modeling decisions after data is collected. This requires a PAP or equivalent pre-registration statement of sufficient detail. Defining what constitutes "sufficient detail" here is an important challenge for the research community to determine, with appropriate account taken of the associated costs and the likely disproportionate burden on particular parts of the research community (see Olken (2015)). The AEA Registry as it currently operates does not place meaningful restrictions on research practices, no doubt allowing for at least some readers to be misled.

Second, if pre-registration is to become a standard some mechanism is needed to ensure greater compliance with pre-registration commitments - that what the researchers end up doing corresponds to what they said they would do. The practice of requiring the PAP be submitted with any journal submission, and passed to referees for inspection at the same time a reading the paper, is an obvious possibility. Again, the costs imposed on reviewers of such extra responsibility must be considered against the benefits. Given high rejection rates at most economics journals such a compliance check could be done through a process similar to the practice of

data verification which is increasingly popular at journals, including those published by the AEA.

Third, given the common recognition that research projects can legitimately and productively evolve as they progress, even after data is collected and is being manipulated, it makes little sense to prohibit publication of results from un-pre-registered elements of work. However it seems prudent that the reader be alerted explicitly to those parts of an article that were not pre-registered, so that they can see them as exploratory in character.

In effect, all three of these suggestions amount to bringing practices in economics closer to those in other disciplines.

Finally, apart from these implications we believe that our data set and analysis makes a descriptive contribution. 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 the period over which the prevalence of pre-registration has grown greatly, from a rarity to comparatively commonplace. By using multiple, state of the art methods to assess the prevalence of p-hacking and publication bias - each with their own strengths and weaknesses, and subject to specific critiques - our investigation provides a more comprehensive picture than would have been the case had we used only a single method.

9 Data Availability

Data and code replicating the tables and figures in this article can be found in Brodeur and Cook (2023) in the Harvard Dataverse, https://doi.org/10.7910/DVN/F8C4YL.

References

Abrams, E., Libgober, J. and List, J.: 2020, Research Registries: Facts, Myths, and Possible Improvements. NBER Working Paper 27250.

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.

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., Carrell, S., Figlio, D. and Lusher, L.: 2023, Unpacking p-Hacking and Publication Bias, *American Economic Review* **113**(11), 2974–3002.

Brodeur, A. and Cook, N.: 2023, Replication Data for: Do Pre-Registration and Pre-Analysis Plans Reduce p-Hacking and Publication Bias?: Evidence from 15,992 Test Statistics and Suggestions for Improvement, *Harvard Dataverse* .

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.: Forthcoming, P-Hacking, Data Type and Availability of Replication Material, *Economic Journal* .

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.: 2023, Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models, *Review of Economic Studies* **90**(1), 201–228.

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.

Gelman, A. and Carlin, J.: 2014, Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors, *Perspectives on Psychological Science* **9**(6), 641–651.

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.

Ioannidis, J. P.: 2008, Why Most Discovered True Associations Are Inflated, *Epidemiology* pp. 640–648.

Ioannidis, J., Stanley, T. and Doucouliagos, H.: 2017, The Power of Bias in Economics Research, *Economic Journal* **127**, F236–F265.

Kranz, S. and Putz, P.: 2022, Methods matter: P-hacking and publication bias in causal analysis in economics: Comment, *American Economic Review* **112**(9), 3124–36.

Kvarven, A., Strømland, E. and Johannesson, M.: 2020, Comparing Meta-Analyses and Preregistered Multiple-Laboratory Replication Projects, *Nature Human Behaviour* **4**(4), 423–434.

Lewis, M., Mathur, M. B., VanderWeele, T. J. and Frank, M. C.: 2022, The Puzzling Relationship Between Multi-Laboratory Replications and Meta-Analyses of the Published Literature, *Royal Society Open Science* **9**(2), 211499.

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

Nosek, B. A., Ebersole, C. R., DeHaven, A. C. and Mellor, D. T.: 2018, The Preregistration Revolution, *Proceedings of the National Academy of Sciences* **115**(11), 2600–2606.

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.

Oostrom, T.: 2022, Funding of Clinical Trials and Reported Drug Efficacy. Ohio State University mimeo.

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.

Simmons, P. J., Nelson, L. D. and Simonsohn, U.: 2021, Pre-Registration: Why and How, *Journal of Consumer Psychology* **31**(1), 151–162.

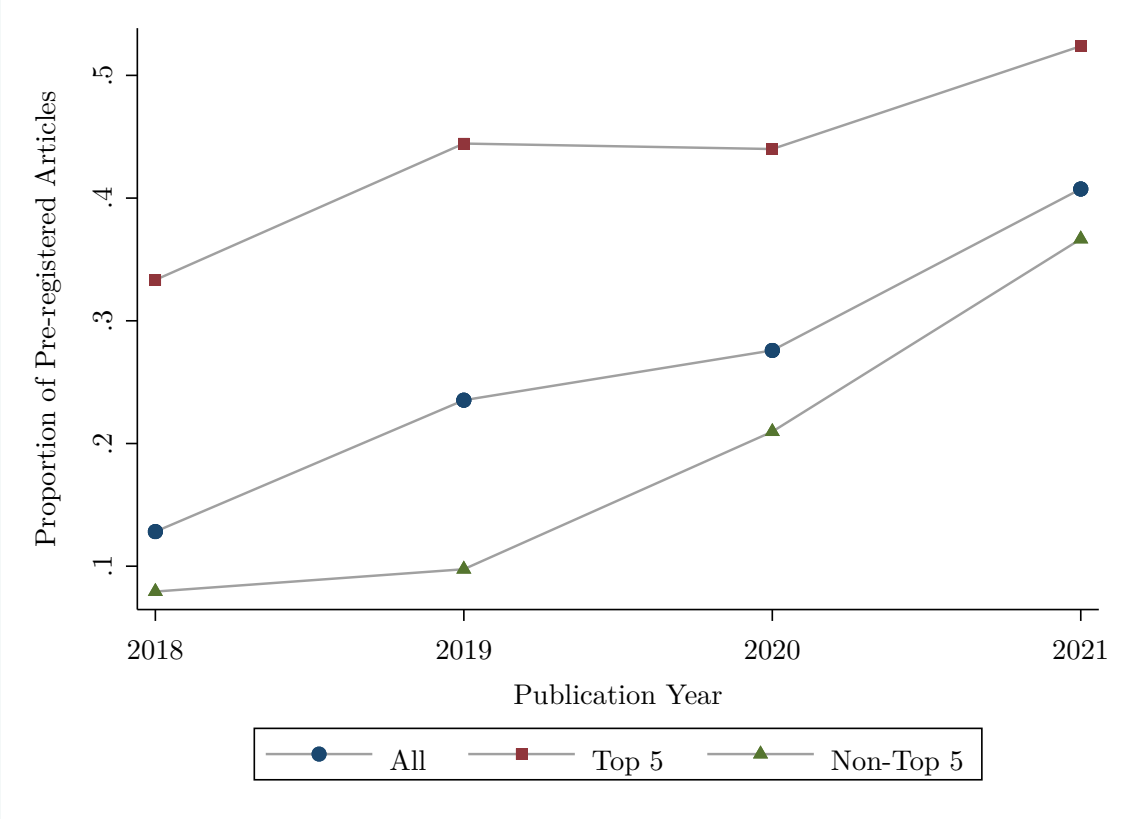
Stanley, T. D. and Doucouliagos, H.: 2014, Meta-Regression Approximations to Reduce Publication Selection Bias, *Research Synthesis Methods* **5**(1), 60–78.

Strømland, E.: 2019, Preregistration and Reproducibility, *Journal of Economic Psychology* **75**, 102143.

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

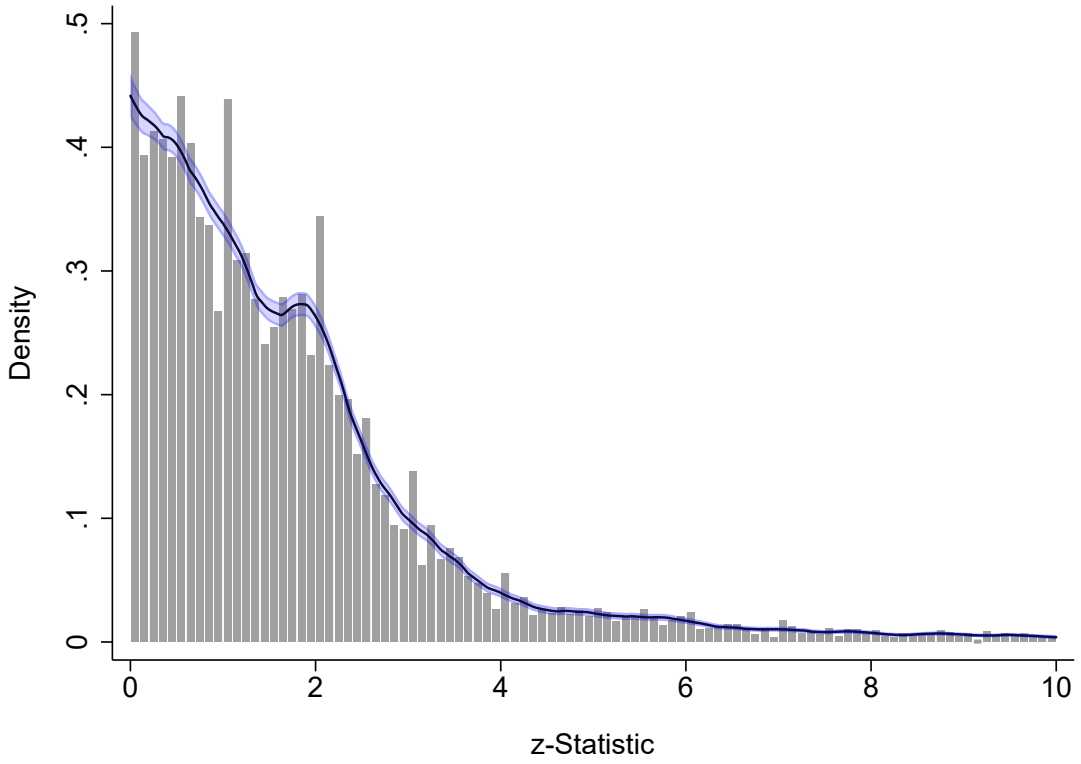
10 Figures

Figure 1: Pre-Registration Rates Over Time



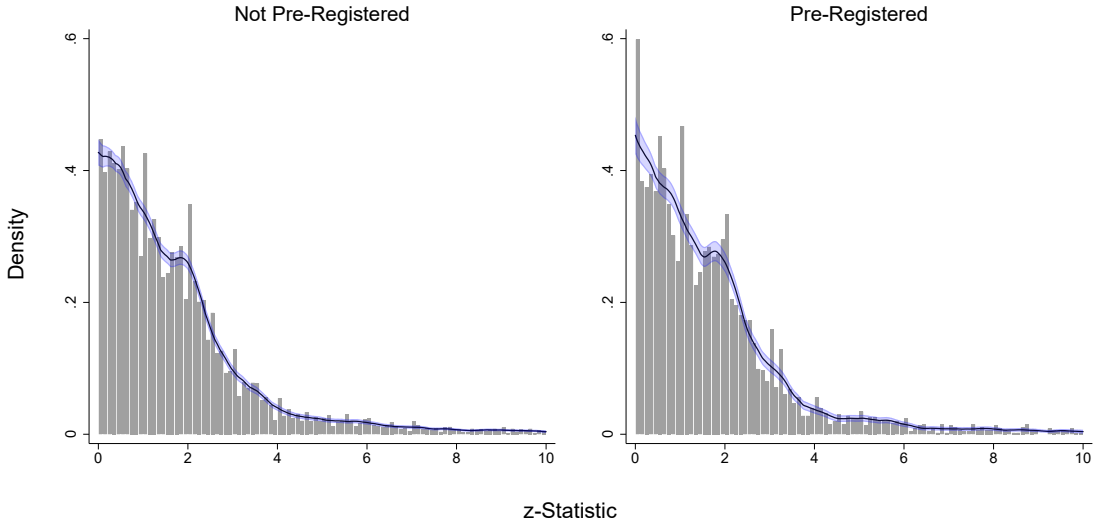
Notes: This figure displays the percentage of pre-registered RCT articles for our full sample and separately for Top 5 and non-Top 5 journals. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were registered after their trial end date are not counted as 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



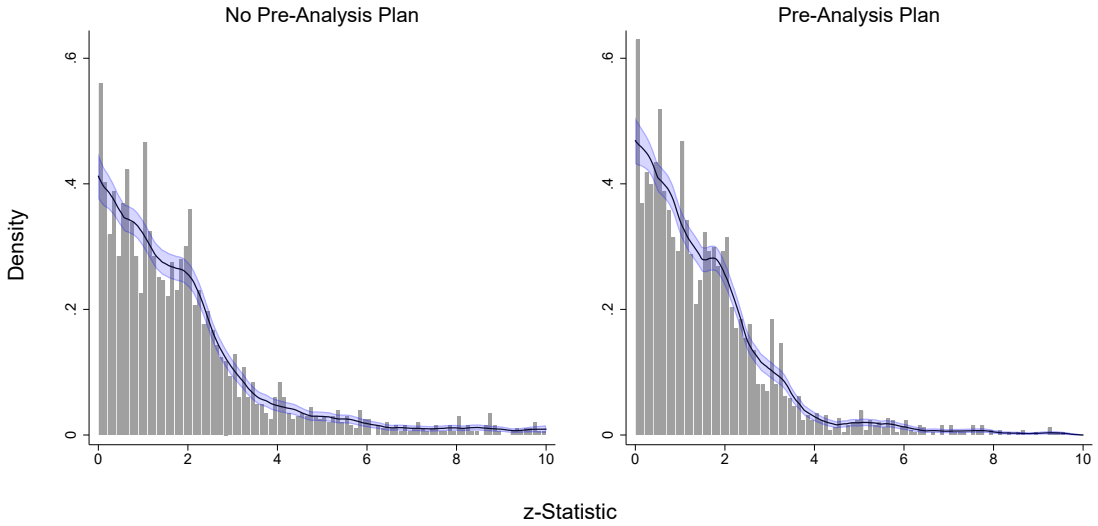
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021. All tests are from articles in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

Figure 3: Test Statistics Distribution by Pre-Registration Status



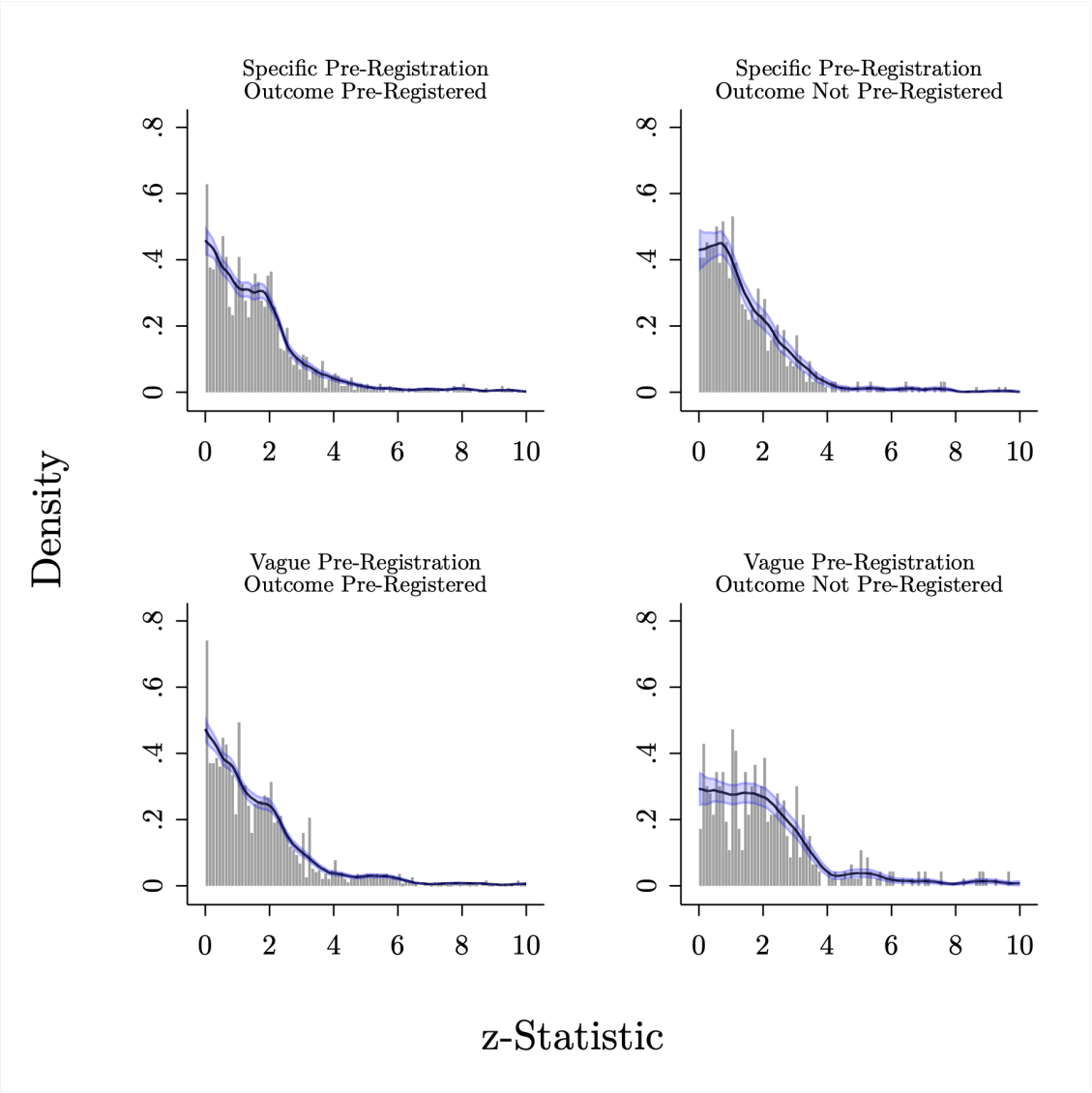
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021 by pre-registration status. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were registered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

Figure 4: Test Statistics Distribution for Pre-Registered RCTs by a Presence of Pre-Analysis Plan



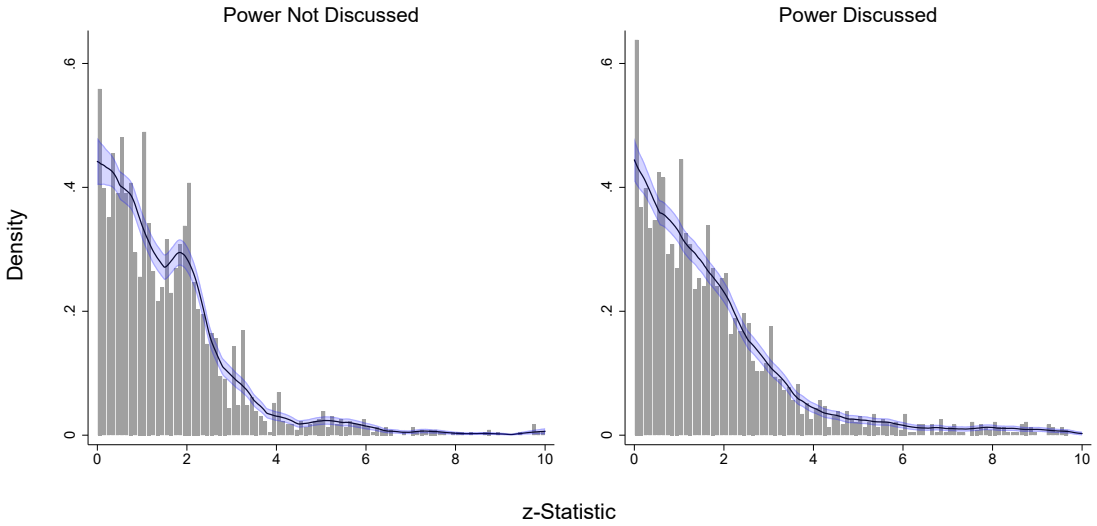
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021 by pre-analysis plan presence. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were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

Figure 5: Test Statistics Distribution by Pre-Registration and Outcome Specificity



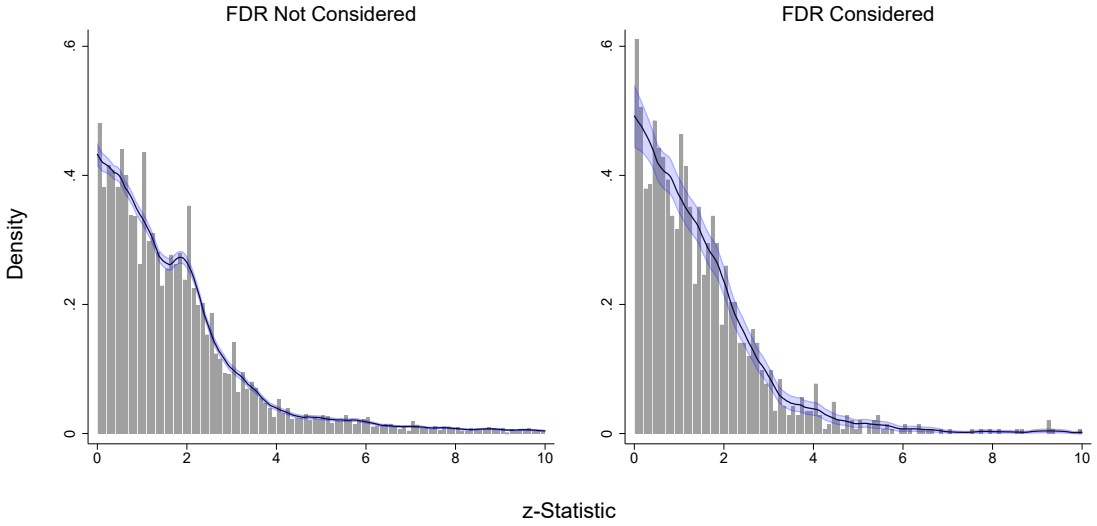
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021 by pre-registration and outcome specificity. Test statistics are divided into those from articles with sufficiently detailed "specific" pre-registrations and while others are more "vague". Test statistics are also divided into those that were pre-specified in the pre-registration and those that were not. Moving from an upper panel to a lower panel represents the movement from a specific to vague pre-registration. Moving from a left panel to a right panel represents moving from a pre-registered test statistic to one not pre-specified. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were registered after their trial end date are not counted as pre-registered. All tests are from articles in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles. For the derounded version see Figure A8.

Figure 6: 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 pre-registered randomized control trials published during 2018–2021 by power analysis status. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were registered after their trial end date are not counted as pre-registered. All tests are from articles 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 False Discovery Rate



Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from pre-registered randomized control trials published during 2018–2021 by the presence of any adjustment for the false discovery rate. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were registered after their trial end date are not counted as pre-registered. All tests are from articles in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

11 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.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 (means). Column 1 includes the full sample of 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

	(1)	(2)	(3)	(4)	(5)	(6)
2019	0.062	0.012	0.034	0.052	-0.027	0.014
	(0.092)	(0.091)	(0.091)	(0.130)	(0.126)	(0.122)
2020	0.146	0.109	0.084	0.129	0.066	0.043
	(0.083)	(0.084)	(0.083)	(0.120)	(0.112)	(0.102)
2021	0.321	0.266	0.261	0.372	0.265	0.259
	(0.094)	(0.088)	(0.086)	(0.129)	(0.112)	(0.109)
Solo Authored	-0.460	-0.443	-0.420	-0.607	-0.609	-0.550
	(0.114)	(0.111)	(0.104)	(0.151)	(0.145)	(0.126)
% Female Authors	-0.235	-0.196	-0.231	-0.369	-0.273	-0.299
	(0.100)	(0.090)	(0.084)	(0.144)	(0.118)	(0.101)
Avg. Experience	0.005	0.002	-0.001	0.033	0.014	-0.001
	(0.022)	(0.020)	(0.019)	(0.035)	(0.027)	(0.025)
Avg. Experience^2	-0.000	-0.000	0.000	-0.001	-0.000	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
PhD Top Institution	0.181	0.136	0.126	0.197	0.137	0.095
	(0.103)	(0.094)	(0.090)	(0.154)	(0.135)	(0.124)
Top Institution	0.120	0.001	0.005	0.154	-0.017	-0.033
	(0.103)	(0.095)	(0.097)	(0.144)	(0.123)	(0.115)
Editor Present	-0.146	-0.156	-0.156	-0.107	-0.140	-0.103
	(0.075)	(0.067)	(0.064)	(0.117)	(0.095)	(0.086)
Top 5		0.253			0.312	
		(0.059)			(0.067)	
AEA Journals		0.015			-0.030	
		(0.066)			(0.081)	
Journal FE			Y			Y
Observations	15,716	15,716	15,563	15,716	15,716	15,563
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	-0.020	-0.029	-0.028	-0.018	-0.028
	(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	3,801	3,710	3,710	2,738	1,634
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: Caliper Test, Statistically Significant at the 10 Percent Level

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	0.014	-0.008	-0.005	0.003	-0.040
	(0.025)	(0.025)	(0.025)	(0.024)	(0.032)
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	4,236	4,146	4,146	2,903	1,592
Window	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]	[1.65±0.35]	[1.65±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 6: 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
Presence Pre-Analysis Plan					
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
Pre-Registration Content					
Pre-Registered, Specific PR	1.706	0.586	0.002	0.000	1.659
Not Pre-Registered, Specific PR	1.370	0.730	0.011	0.006	1.516
Pre-Registered, Vague PR	1.773	0.564	0.005	0.004	1.503
Not Pre-Registered, Vague PR	1.898	0.527	0.055	0.041	1.988
Power Discussion					
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
False Discovery Rate Considered					
Mention of FDR	1.182	0.846	0.025	0.021	1.757
No Mention of FDR	1.712	0.584	0.013	0.012	1.712

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 7: Caliper Test, Statistically Significant at the 5 Percent Level: Completeness of Pre-Analysis Plan

	(1)	(2)	(3)	(4)	(5)
Pre-Analysis Plan	-0.071	-0.106	-0.122	-0.149	-0.122
	(0.044)	(0.043)	(0.040)	(0.047)	(0.048)
Power Discussed	-0.039	-0.000	0.007	-0.026	0.032
	(0.052)	(0.051)	(0.050)	(0.059)	(0.063)
Specific	-0.038	-0.076	-0.027	-0.016	-0.093
	(0.072)	(0.067)	(0.069)	(0.080)	(0.084)
Pre-Registered Test	-0.057	-0.051	-0.022	-0.017	0.029
	(0.069)	(0.062)	(0.061)	(0.060)	(0.067)
Specific * Pre-Reg	0.071	0.051	0.049	0.114	0.157
	(0.094)	(0.089)	(0.090)	(0.094)	(0.100)
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.

Table 8: Caliper Test, Statistically Significant at the 10 Percent Level: Completeness of Pre-Analysis Plan

	(1)	(2)	(3)	(4)	(5)
Pre-Analysis Plan	0.008	0.013	-0.012	0.011	-0.001
	(0.042)	(0.036)	(0.035)	(0.042)	(0.051)
Power Discussed	-0.053	-0.017	-0.030	-0.020	0.056
	(0.047)	(0.044)	(0.041)	(0.051)	(0.059)
Specific	-0.039	-0.040	0.008	0.035	0.190
	(0.067)	(0.066)	(0.067)	(0.085)	(0.123)
Pre-Registered Test	-0.019	0.014	0.039	0.067	0.159
	(0.063)	(0.069)	(0.065)	(0.074)	(0.114)
Specific * Pre-Reg	0.042	0.017	-0.008	-0.093	-0.269
	(0.085)	(0.083)	(0.078)	(0.101)	(0.138)
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	1,278	1,253	1,249	879	495
Window	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]	[1.65±0.35]	[1.65±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.

12 ONLINE APPENDIX

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

We rely on 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 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)[36] 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 Appendix Figure A11, 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

[36]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.

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.

12.2 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 Appendix Figure A12 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 Appendix Figure A12. 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 (2023). 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

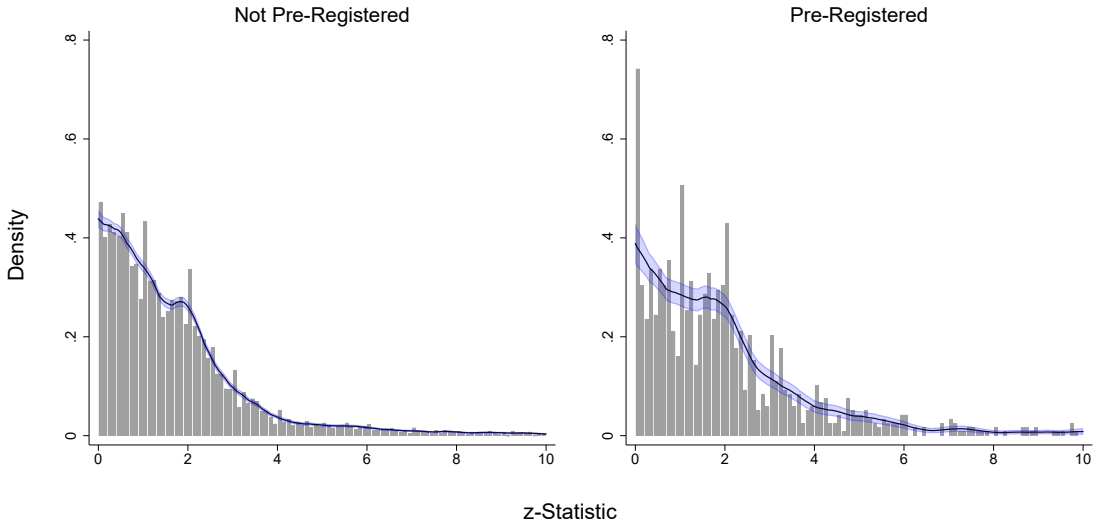
histogram based test designed against its 2-monotonicity and places bounds on the p-curve and its first two derivatives. Fifth, and our first major difference between the samples, a test whose null hypothesis is that the p-value distribution (not the p-curve) is concave. Applying this Least Concave Majorant (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.

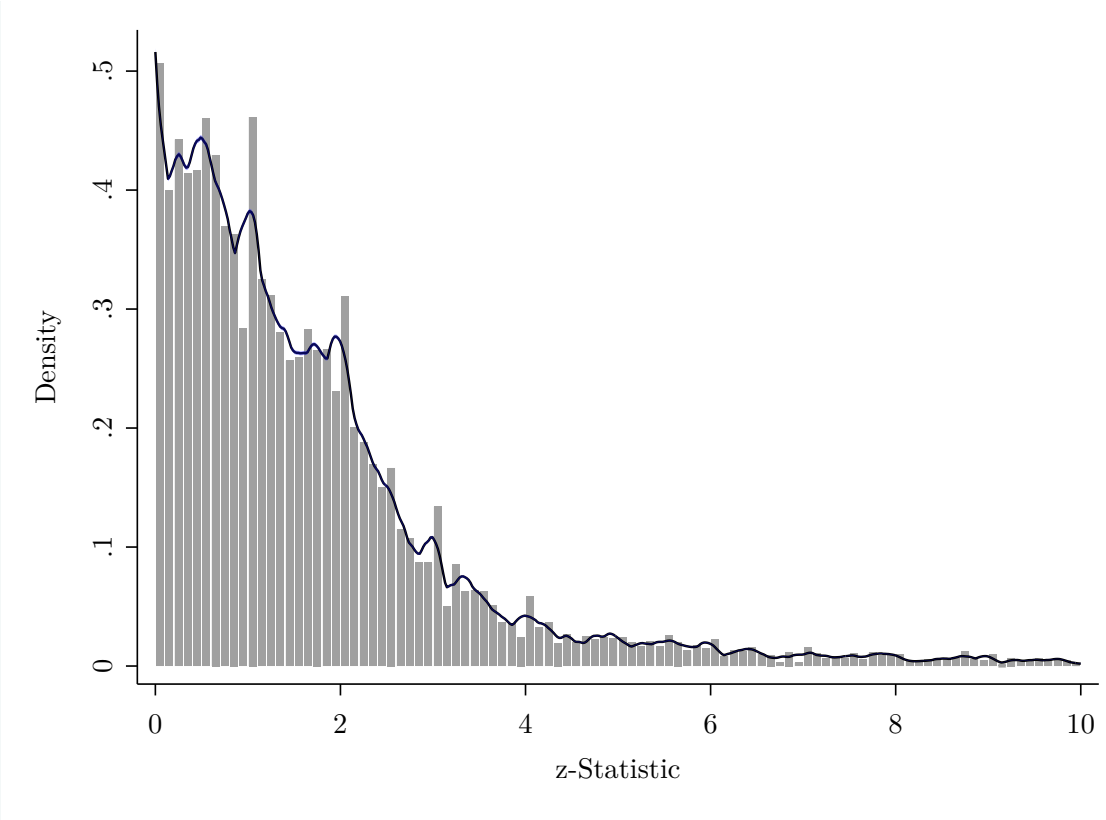
13 Appendix Figures

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



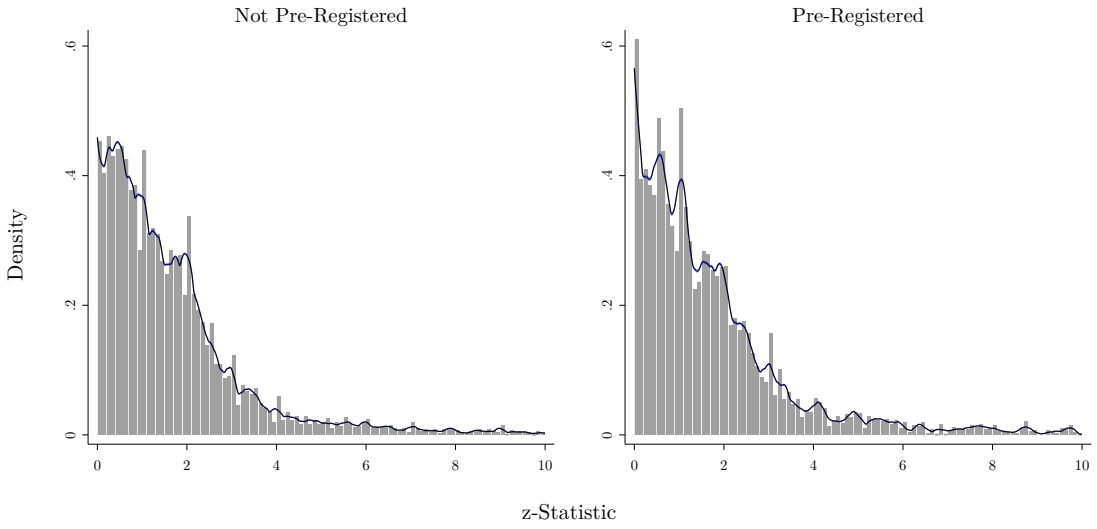
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021 by pre-registration status. Here we *alternatively* define a pre-registered RCT as a study that was registered on a date prior to its registry trial *start* date. Studies that were registered after their trial *start* date are not counted as pre-registered. All tests are from articles 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



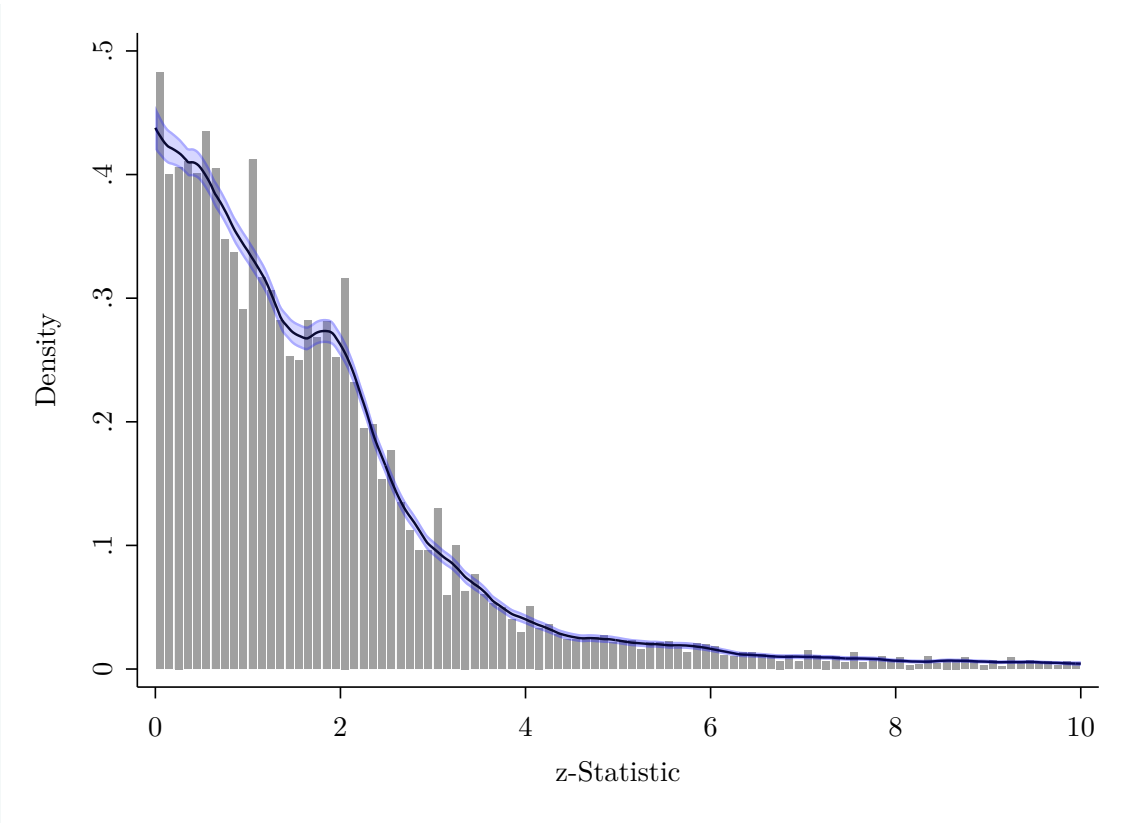
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021. All tests are from articles in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We use the inverse of the number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

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



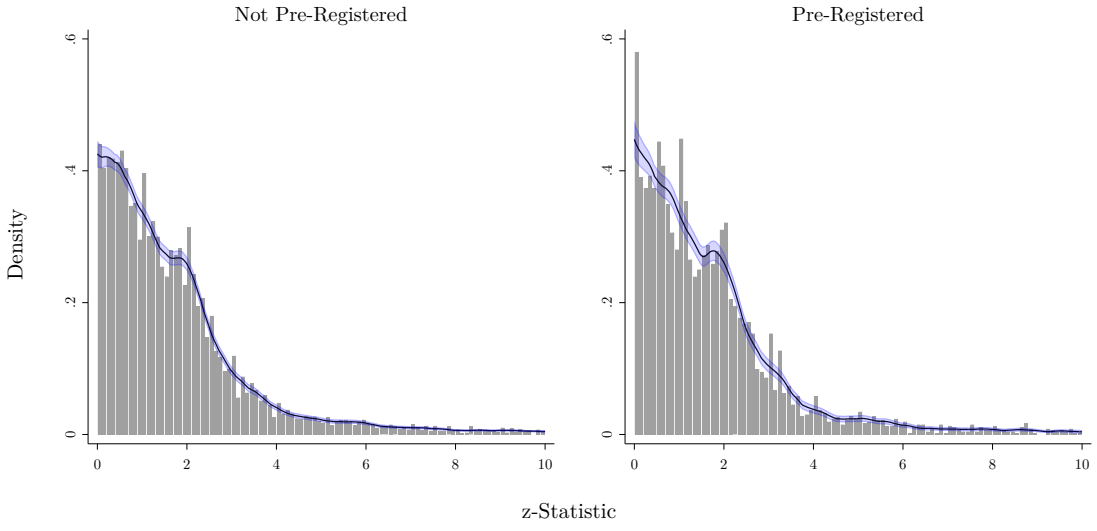
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021 by pre-registration status. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were registered after their trial end date are not counted as pre-registered. All tests are from articles 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



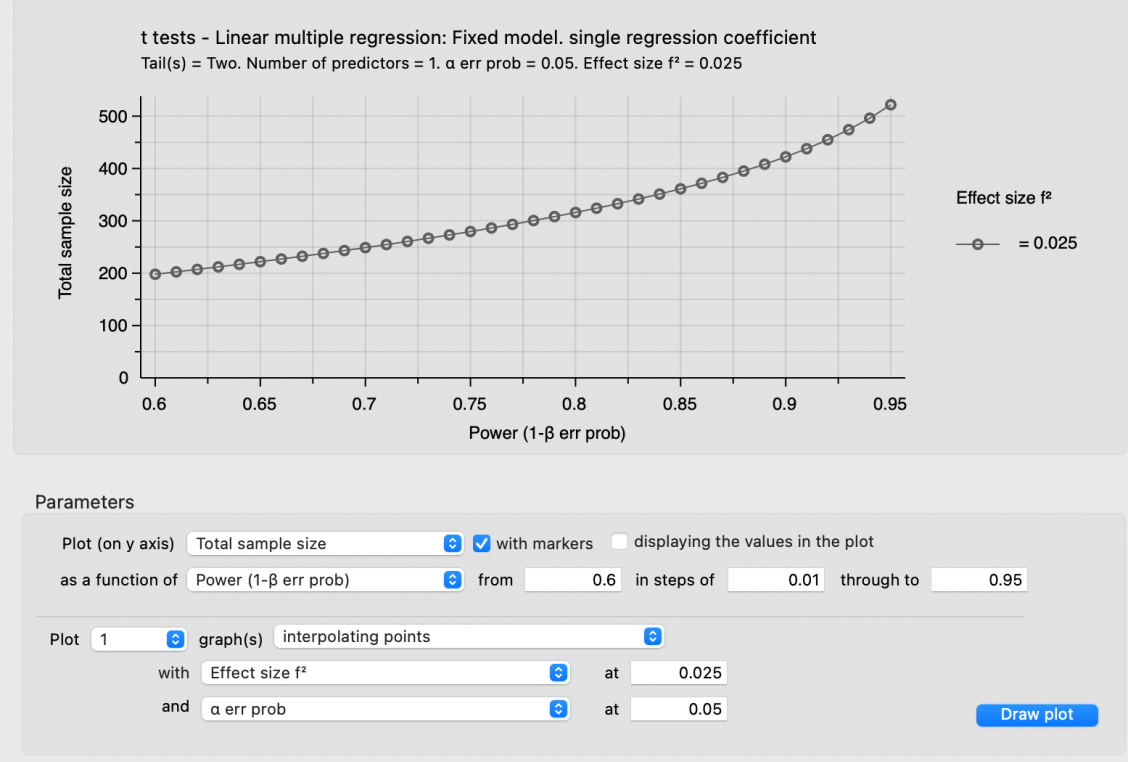
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021. All tests are from articles in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles. We apply test statistics derounding following Brodeur et al. (2016).

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



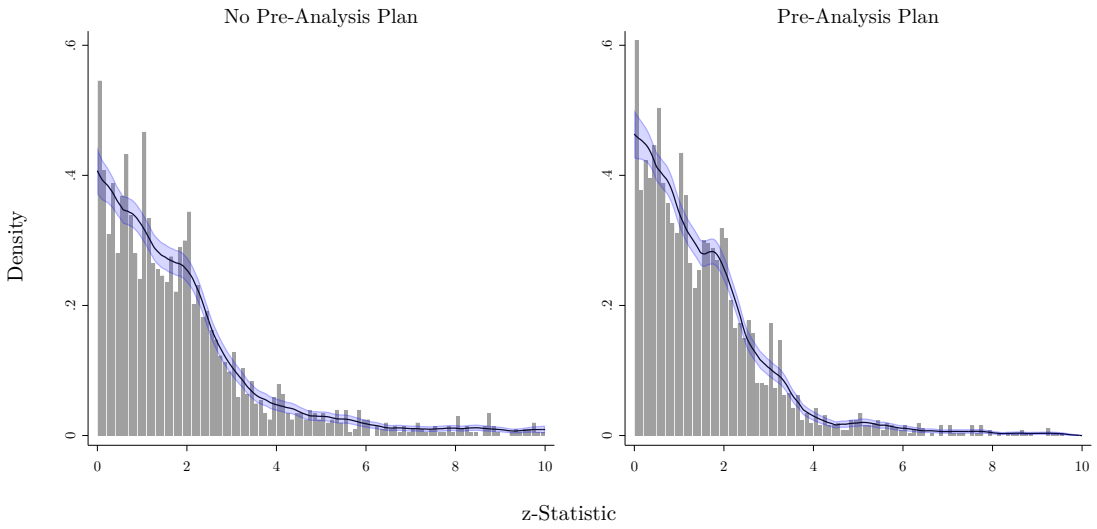
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021 by pre-registration status. All tests are from articles in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles. We apply test statistics derounding following Brodeur et al. (2016).

Figure A6: A Priori Power Test



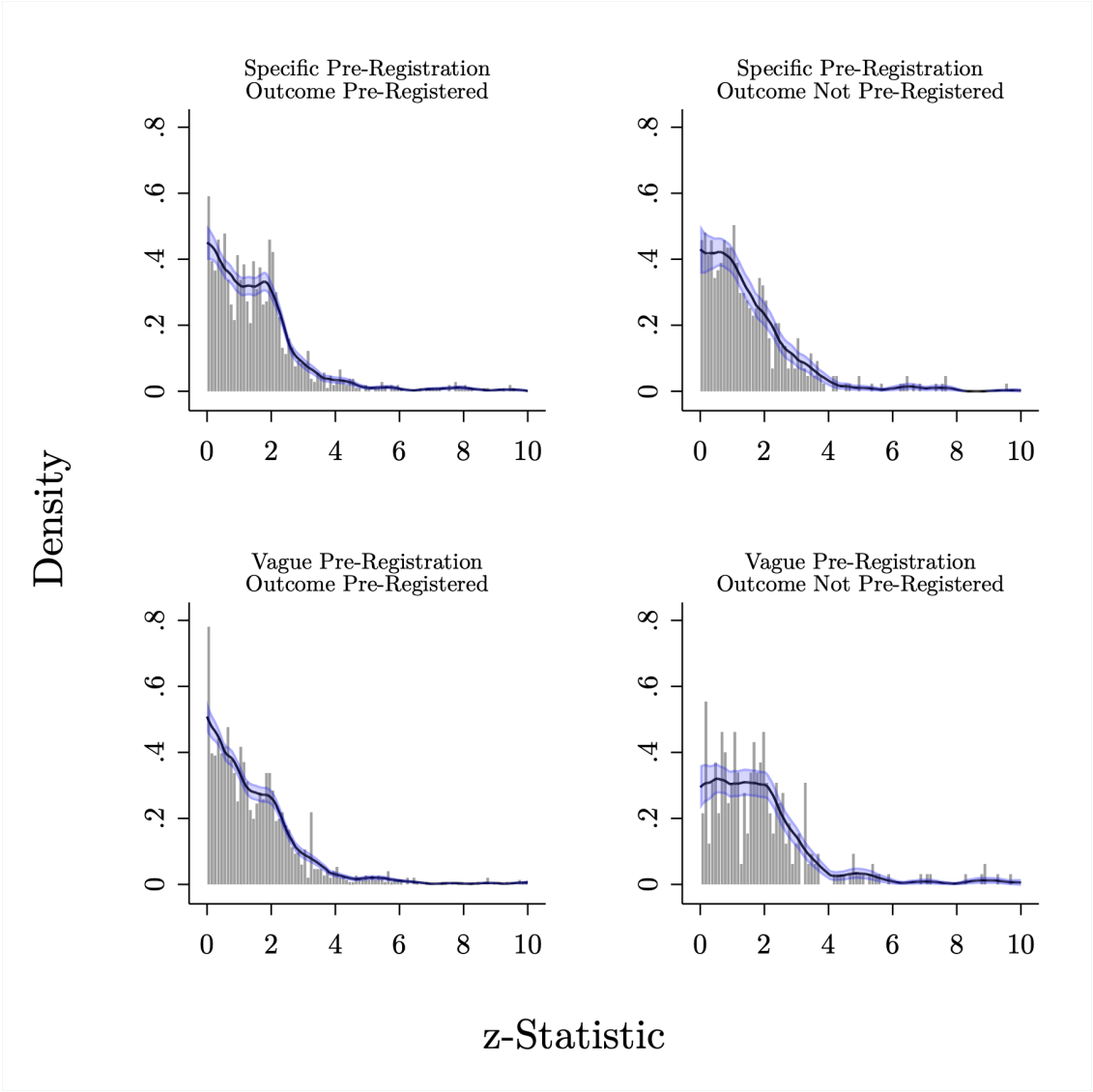
Notes: This figure shows our a-priori statistical power calculations. We wished to detect a small effect with $f^2 = 0.025$. We expected that our sample size for the caliper test using $z \in [1.46, 2.46]$ would be over 6,000. We also expected that approximately 33% of test statistics would be in pre-registered RCTs.

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



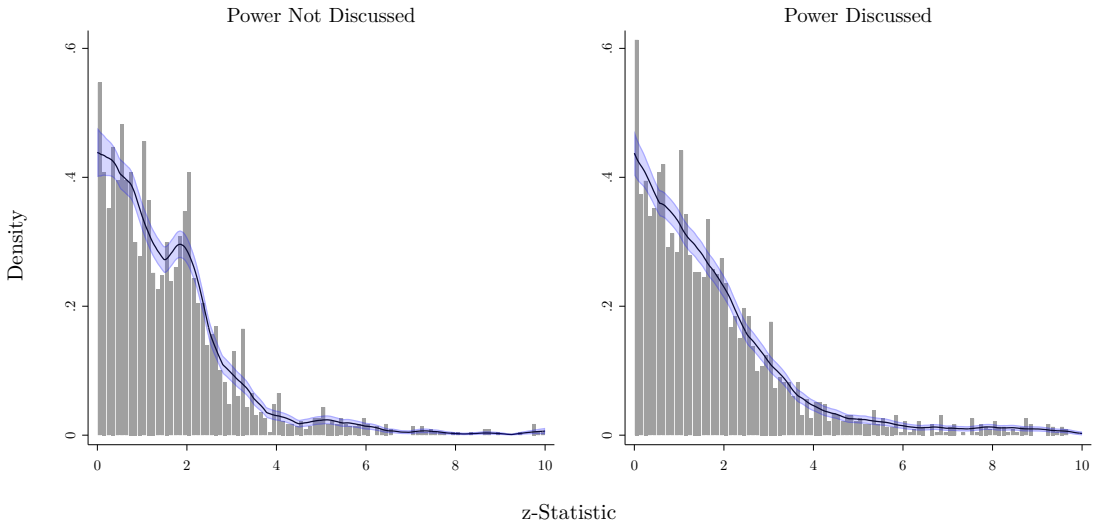
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021 by pre-analysis plan presence. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were registered after their trial end date are not counted as pre-registered. All tests are from articles in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles. We apply test statistics derounding following Brodeur et al. (2016).

Figure A8: Test Statistics Distribution by Pre-Registration and Outcome Specificity: Derounding



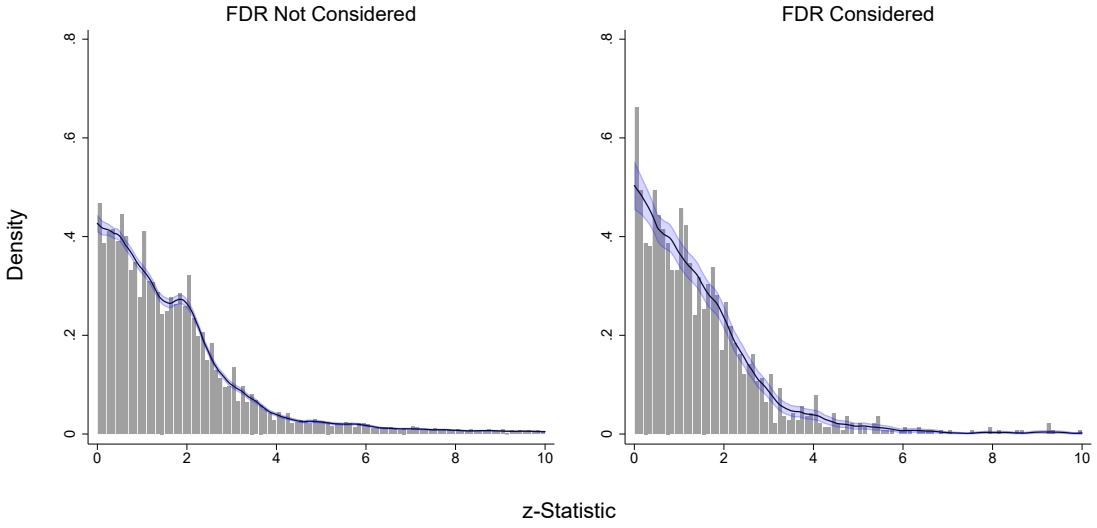
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021 by pre-registration and outcome specificity. Test statistics are derounded following Kranz and Putz (2022). Test statistics are divided into those from articles with sufficiently detailed "specific" pre-registrations and while others are more "vague". Test statistics are also divided into those that were pre-specified in the pre-registration and those that were not. Moving from an upper panel to a lower panel represents the movement from a specific to vague pre-registration. Moving from a left panel to a right panel represents moving from a pre-registered test statistic to one not pre-specified. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were registered after their trial end date are not counted as pre-registered. All tests are from articles in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

Figure A9: Test Statistics Distribution for Pre-Registered RCTs by Presence of Power Analysis: Derounding



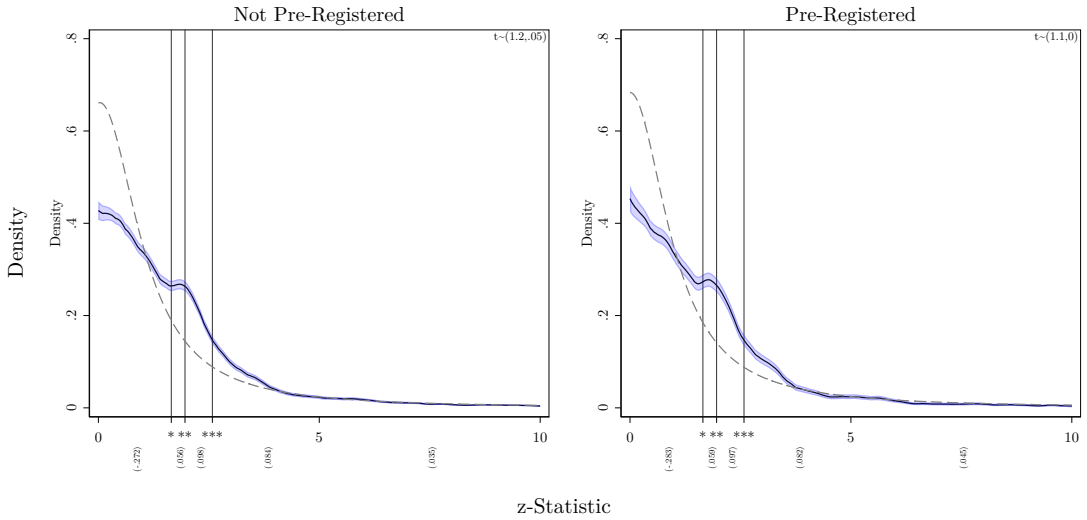
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from pre-registered randomized control trials published during 2018–2021 by power analysis status. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were registered after their trial end date are not counted as pre-registered. All tests are from articles in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles. We apply test statistics derounding following Brodeur et al. (2016).

Figure A10: Test Statistics Distribution for Pre-Registered RCTs by False Discovery Rate - q-Value: Derounding



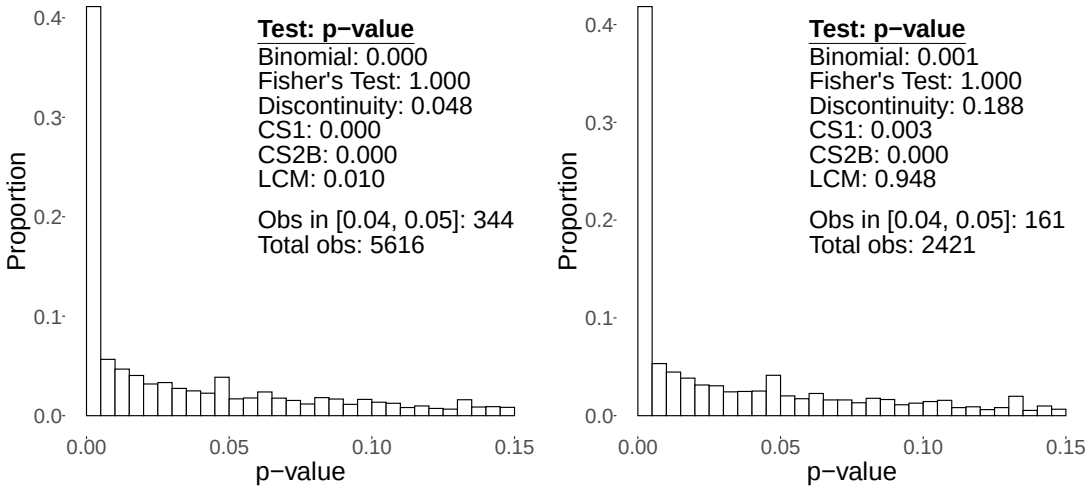
Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from pre-registered randomized control trials published during 2018–2021 by the presence of any adjustment for the false discovery rate. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were registered after their trial end date are not counted as pre-registered. All tests are from articles in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles. We apply test statistics derounding following Brodeur et al. (2016).

Figure A11: Excess Test Statistics by Pre-Registration Status



Notes: This figure displays the observed distribution of test statistics for $z \in [0, 10]$ from randomized control trials published during 2018–2021 by pre-registration status. Also displayed are status-specific counterfactual distributions (dashed lines) we would 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 observed-expected mass difference between statistical significance thresholds. We define a pre-registered RCT as a study that was registered on a date prior to its registry trial end date. Studies that were registered after their trial end date are not counted as pre-registered. All tests are from articles in 15 leading economics journals. We do not weight articles.

Figure A12: Application of Elliott et al. (2022) by Pre-Registration Status



Notes: Each panel displays a direct application of Elliott et al. (2022)'s p-hacking detecting test to subsamples defined by pre-registration status. The left panel displays a p-curve for a pre-registered RCT's (registered on a date prior to registry trial end date). The right panel displays a p-curve for non-pre-registered RCT's (registered on a date after registry trial end date). All tests are from articles in 15 leading economics journals. Section 12.2 discusses the p-curve and included tests in detail.

14 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. In column 1, we report the number of articles from each journal. In column 2, we report the number of tests from each journal. In column 3, we report the proportion of articles that have a pre-registration.

Table A2: Prediction of Pre-Registration Use (Linear Probability Models)

	(1)	(2)	(3)	(4)	(5)	(6)
2019	0.054	0.004	0.034	0.037	-0.028	0.014
	(0.096)	(0.091)	(0.091)	(0.137)	(0.132)	(0.132)
2020	0.139	0.104	0.079	0.111	0.058	0.033
	(0.082)	(0.082)	(0.084)	(0.118)	(0.113)	(0.108)
2021	0.315	0.266	0.257	0.350	0.253	0.245
	(0.091)	(0.086)	(0.086)	(0.127)	(0.111)	(0.114)
Solo Authored	-0.368	-0.366	-0.357	-0.305	-0.316	-0.258
	(0.086)	(0.088)	(0.095)	(0.142)	(0.141)	(0.147)
% Female Authors	-0.220	-0.174	-0.211	-0.338	-0.235	-0.270
	(0.091)	(0.084)	(0.082)	(0.138)	(0.115)	(0.108)
Avg. Experience	0.007	0.004	0.002	0.033	0.016	0.001
	(0.023)	(0.022)	(0.023)	(0.037)	(0.032)	(0.032)
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)	(0.001)
PhD Top Institution	0.191	0.132	0.119	0.214	0.128	0.094
	(0.102)	(0.093)	(0.091)	(0.155)	(0.139)	(0.132)
Top Institution	0.105	0.000	0.005	0.112	-0.015	-0.041
	(0.100)	(0.090)	(0.089)	(0.132)	(0.114)	(0.111)
Editor Present	-0.158	-0.176	-0.175	-0.115	-0.173	-0.136
	(0.084)	(0.076)	(0.072)	(0.134)	(0.109)	(0.099)
Top 5		0.288			0.364	
		(0.075)			(0.100)	
AEA Journals		0.042			0.019	
		(0.070)			(0.092)	
Constant	0.240	0.232	0.271	0.101	0.192	0.231
	(0.163)	(0.152)	(0.173)	(0.239)	(0.213)	(0.224)
Journal FE			Y			Y
Observations	15,716	15,716	15,716	15,716	15,716	15,716
Weights				Article	Article	Article

Notes: This table reports a linear probability model. 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 A3: Caliper Test, Statistically Significant at the 5 Percent Level: Logit

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	-0.020	-0.029	-0.028	-0.018	-0.028
	(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	3,801	3,710	3,710	2,738	1,634
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 A4: Caliper Test, Statistically Significant at the 10 Percent Level (Logit)

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	0.014	-0.008	-0.005	0.003	-0.040
	(0.025)	(0.025)	(0.025)	(0.024)	(0.032)
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Article's Charact.		Y	Y	Y	Y
AEA and Top 5		Y			
Journal FE			Y	Y	Y
Observations	4,236	4,146	4,146	2,903	1,592
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 10 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 (Linear)

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	-0.020	-0.029	-0.028	-0.018	-0.028
	(0.027)	(0.026)	(0.027)	(0.029)	(0.036)
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Article's Charact.		Y	Y	Y	Y
AEA and Top 5		Y			
Journal FE			Y	Y	Y
Observations	3,801	3,710	3,710	2,738	1,634
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 a linear probability model. 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 A6: Caliper Test, Statistically Significant at the 10 Percent Level (Linear)

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	0.014	-0.008	-0.005	0.003	-0.040
	(0.025)	(0.025)	(0.025)	(0.025)	(0.032)
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Article's Charact.		Y	Y	Y	Y
AEA and Top 5		Y			
Journal FE			Y	Y	Y
Observations	4,236	4,146	4,146	2,903	1,593
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 a linear probability model. 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 A7: Caliper Test, Statistically Significant at the 5 Percent Level: Derounding

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	-0.016	-0.028	-0.025	-0.019	-0.020
	(0.027)	(0.026)	(0.027)	(0.029)	(0.034)
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	3,783	3,692	3,692	2,727	1,626
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. Derounding applied following Brodeur et al. (2016).

Table A8: Caliper Test, Statistically Significant at the 10 Percent Level (Derounding)

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	0.027	0.004	0.009	0.021	-0.030
	(0.025)	(0.025)	(0.025)	(0.024)	(0.029)
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Article's Charact.		Y	Y	Y	Y
AEA and Top 5		Y			
Journal FE			Y	Y	Y
Observations	4,237	4,147	4,147	2,897	1,596
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. Derounding applied following Brodeur et al. (2016).

Table A9: Caliper Test, Statistically Significant at the 1 Percent Level

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	0.009	0.011	0.004	-0.034	-0.025
	(0.033)	(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	2,363	2,321	2,321	1,536	879
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 A10: Caliper Test, Statistically Significant at the 5 Percent Level: Article Weights

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	-0.031	-0.043	-0.044	-0.022	-0.032
	(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	3,801	3,710	3,710	2,738	1,634
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. We use the inverse of the number of tests presented in the same article to weight observations.

Table A11: Prediction of Pre-Analysis Plan Use (Only Pre-Registered)

	(1)	(2)	(3)	(4)	(5)	(6)
2019	0.197	0.138	0.193	0.287	0.243	0.441
	(0.227)	(0.219)	(0.168)	(0.260)	(0.259)	(0.183)
2020	0.127	0.136	0.040	0.284	0.352	0.147
	(0.255)	(0.229)	(0.198)	(0.301)	(0.250)	(0.240)
2021	-0.064	-0.046	0.060	-0.022	0.056	0.167
	(0.228)	(0.214)	(0.172)	(0.264)	(0.238)	(0.177)
% Female Authors	0.149	0.168	0.371	0.077	0.150	0.527
	(0.230)	(0.221)	(0.207)	(0.249)	(0.244)	(0.203)
Avg. Experience	0.002	-0.025	-0.008	0.023	-0.027	-0.101
	(0.055)	(0.053)	(0.054)	(0.081)	(0.079)	(0.097)
Avg. Experience^2	0.000	0.002	0.001	-0.001	0.002	0.005
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.004)
PhD Top Institution	-0.524	-0.533	-0.448	-0.644	-0.569	-0.528
	(0.206)	(0.187)	(0.178)	(0.261)	(0.231)	(0.181)
Top Institution	0.407	0.196	0.100	0.713	0.351	0.176
	(0.264)	(0.271)	(0.265)	(0.258)	(0.319)	(0.265)
Editor Present	0.039	0.060	-0.101	0.071	0.134	-0.247
	(0.173)	(0.157)	(0.163)	(0.239)	(0.216)	(0.204)
Top 5		0.256			0.269	
		(0.146)			(0.181)	
AEA Journals		-0.177			-0.278	
		(0.134)			(0.155)	
Journal FE			Y			Y
Observations	4,766	4,766	4,439	4,766	4,766	4,439

Notes: This table reports marginal effects from probit regressions. The sample is restricted to only observations from studies that have a pre-registration. The dependent variable is a dummy for the presence of pre-analysis plan. 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 A12: Prediction of Pre-Analysis Plan Use

	(1)	(2)	(3)	(4)	(5)	(6)
2019	0.018	-0.011	0.009	0.015	-0.033	0.006
	(0.086)	(0.083)	(0.081)	(0.128)	(0.120)	(0.107)
2020	0.069	0.044	0.059	0.075	0.037	0.045
	(0.077)	(0.075)	(0.069)	(0.111)	(0.108)	(0.084)
2021	0.146	0.111	0.145	0.140	0.072	0.123
	(0.088)	(0.086)	(0.082)	(0.133)	(0.126)	(0.113)
Solo Authored	-0.016	0.002	-0.023	0.112	0.133	0.062
	(0.146)	(0.142)	(0.139)	(0.229)	(0.216)	(0.201)
% Female Authors	0.025	0.045	0.033	-0.075	-0.012	0.031
	(0.097)	(0.094)	(0.094)	(0.146)	(0.136)	(0.128)
Avg. Experience	0.010	0.004	-0.003	0.050	0.028	0.018
	(0.023)	(0.021)	(0.020)	(0.041)	(0.034)	(0.030)
Avg. Experience^2	-0.000	-0.000	0.000	-0.002	-0.001	-0.000
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
PhD Top Institution	0.013	-0.012	-0.016	0.021	-0.016	-0.032
	(0.084)	(0.079)	(0.081)	(0.122)	(0.112)	(0.106)
Top Institution	0.060	-0.044	-0.097	0.108	-0.044	-0.177
	(0.102)	(0.095)	(0.099)	(0.149)	(0.135)	(0.137)
Editor Present	-0.073	-0.076	-0.072	-0.096	-0.123	-0.104
	(0.072)	(0.067)	(0.065)	(0.113)	(0.098)	(0.087)
Top 5		0.198			0.268	
		(0.061)			(0.082)	
AEA Journals		-0.034			-0.071	
		(0.065)			(0.085)	
Journal FE			Y			Y
Observations	15,716	15,716	15,050	15,716	15,716	15,050

Notes: This table reports marginal effects from probit regressions. The dependent variable is a dummy for the presence of pre-analysis plan. 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 A13: Caliper Test, Statistically Significant at the 5 Percent Level: Presence of 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.

Table A14: Caliper Test, Statistically Significant at the 5 Percent Level: Completeness of Pre-Analysis Plan (Derounding)

	(1)	(2)	(3)	(4)	(5)
Pre-Analysis Plan	-0.080	-0.115	-0.132	-0.158	-0.157
	(0.043)	(0.042)	(0.038)	(0.045)	(0.045)
Power Discussed	-0.041	0.003	0.007	-0.019	0.024
	(0.052)	(0.051)	(0.051)	(0.060)	(0.060)
Specific	-0.022	-0.069	-0.024	-0.017	-0.089
	(0.072)	(0.065)	(0.066)	(0.076)	(0.078)
Pre-Registered Test	-0.046	-0.043	-0.019	-0.026	-0.025
	(0.068)	(0.065)	(0.065)	(0.067)	(0.068)
Specifc × Pre-Reg.	0.061	0.047	0.050	0.114	0.175
	(0.094)	(0.088)	(0.091)	(0.098)	(0.101)
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	839	515
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. Derounding applied following Brodeur et al. (2016).

Table A15: Caliper Test, Statistically Significant at the 5 Percent Level: Completeness of Pre-Analysis Plan, Full Sample

	(1)	(2)	(3)	(4)	(5)
Pre-Registered Only	0.015	0.015	0.012	0.031	0.011
	(0.031)	(0.036)	(0.037)	(0.039)	(0.044)
PR and Pre-Analysis Plan	-0.048	-0.013	0.000	-0.041	-0.046
	(0.037)	(0.085)	(0.089)	(0.093)	(0.083)
Power Discussed		-0.086	-0.094	-0.165	-0.178
		(0.063)	(0.063)	(0.068)	(0.082)
Specific		-0.001	-0.007	0.083	0.048
		(0.089)	(0.094)	(0.106)	(0.109)
Pre-Registered Test		-0.073	-0.080	-0.066	-0.114
		(0.086)	(0.084)	(0.093)	(0.069)
Specific × Pre-Reg.		0.091	0.098	0.123	0.236
		(0.118)	(0.118)	(0.129)	(0.123)
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Article's Charact.		Y	Y	Y	Y
AEA and Top 5		Y			
Journal FE			Y	Y	Y
Observations	3,801	3,710	3,710	2,738	1,634
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. We include all RCTs in our sample. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table A16: Caliper Test, Statistically Significant at the 10 Percent Level: Completeness of Pre-Analysis Plan, Full Sample

	(1)	(2)	(3)	(4)	(5)
Pre-Registered Only	0.012	-0.013	-0.003	0.006	-0.034
	(0.036)	(0.038)	(0.038)	(0.041)	(0.049)
PR and Pre-Analysis Plan	0.016	0.053	0.056	0.055	-0.030
	(0.030)	(0.069)	(0.070)	(0.079)	(0.114)
Power Discussed		-0.046	-0.073	-0.001	0.090
		(0.062)	(0.062)	(0.057)	(0.067)
Specific		-0.123	-0.111	-0.118	-0.030
		(0.081)	(0.083)	(0.107)	(0.141)
Pre-Registered Test		-0.077	-0.064	-0.024	0.025
		(0.094)	(0.094)	(0.106)	(0.162)
Specifc × Pre-Reg.		0.214	0.195	0.084	-0.115
		(0.108)	(0.109)	(0.135)	(0.187)
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Article's Charact.		Y	Y	Y	Y
AEA and Top 5		Y			
Journal FE			Y	Y	Y
Observations	4,236	4,146	4,146	2,903	1,592
Window	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]	[1.65±0.35]	[1.65±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. We include all RCTs in our sample. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table A17: Caliper Test, Statistically Significant at the 5 Percent Level: Completeness of Pre-Analysis Plan (Linear)

	(1)	(2)	(3)	(4)	(5)
Pre-Analysis Plan	-0.071	-0.107	-0.123	-0.149	-0.122
	(0.044)	(0.044)	(0.041)	(0.049)	(0.052)
Power Discussed	-0.039	-0.000	0.008	-0.024	0.039
	(0.053)	(0.052)	(0.052)	(0.062)	(0.066)
Specific	-0.038	-0.079	-0.031	-0.022	-0.105
	(0.073)	(0.069)	(0.070)	(0.083)	(0.090)
Pre-Registered Test	-0.057	-0.053	-0.025	-0.018	0.028
	(0.069)	(0.063)	(0.063)	(0.061)	(0.070)
Specific × Pre-Reg.	0.071	0.055	0.052	0.114	0.158
	(0.094)	(0.090)	(0.092)	(0.097)	(0.103)
Reporting Method		Y	Y	Y	Y
Authors' Charact.		Y	Y	Y	Y
Article's Charact.		Y	Y	Y	Y
AEA and Top 5		Y			
Journal FE			Y	Y	Y
Observations	1,164	1,135	1,135	839	511
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 a linear probability model. 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 A18: Caliper Test, Statistically Significant at the 5 Percent Level: False Discovery Rate

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	-0.015	-0.032	-0.031	-0.021	-0.033
	(0.027)	(0.027)	(0.027)	(0.030)	(0.036)
False Discovery (q-value)	-0.068	-0.052	-0.060	-0.035	-0.054
	(0.040)	(0.039)	(0.038)	(0.043)	(0.059)
Controls					
Reporting Method					
Authors' Charact.		Y	Y	Y	Y
Articles' Charact.		Y	Y	Y	Y
AEA & Top 5		Y			
Journal FE			Y	Y	Y
Observations	3,801	3,710	3,710	2,738	1,634
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. The control variables are similar to Table 4 with the exception that we exclude reporting methods. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table A19: Caliper Test, Statistically Significant at the 10 Percent Level: False Discovery Rate

	(1)	(2)	(3)	(4)	(5)
Pre-Registered	0.018	-0.005	-0.002	0.003	-0.043
	(0.025)	(0.025)	(0.025)	(0.025)	(0.032)
False Discovery (q-value)	-0.046	-0.038	-0.042	0.004	0.051
	(0.026)	(0.027)	(0.027)	(0.031)	(0.043)
Controls					
Reporting Method					
Authors' Charact.		Y	Y	Y	Y
Articles' Charact.		Y	Y	Y	Y
AEA & Top 5		Y			
Journal FE			Y	Y	Y
Observations	4,236	4,146	4,146	2,903	1,592
Window	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]	[1.65±0.35]	[1.65±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.