ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 200

P-Hacking, Data Type and Data-Sharing Policy

Abel Brodeur Nikolai Cook Carina Neisser

September 2022 www.econtribute.de

UNIVERSITÄT BONN Universität zu Köln

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866 is gratefully acknowledged.

P-Hacking, Data Type and Data-Sharing Policy*

Abel Brodeur Nikolai Cook Carina Neisser

September 26, 2022

Abstract

In this paper, we examine the relationship between p-hacking and data-sharing policies for published articles. We collect 38,876 test statistics from 1,106 articles published in leading economic journals between 2002–2020. While a data-sharing policy increases the provision of research data to the community, we find a well-estimated null effect that requiring authors to share their data at the time of publication does not alter the presence of p-hacking. Similarly, articles that use hard-to-access administrative data or third-party surveys, as compared to those that use easier-to-access (e.g., own-collected) data are not different in their p-hacking extent. Voluntary provision of data by authors on their homepages offers no evidence of reduced p-hacking.

KEYWORDS: p-Hacking - Publication Bias - Data and Code Availability - Data Sharing Policy - Administrative Data - Survey Data

JEL CODES: A11, B41, C13, C40, I23.

*Authors: Brodeur: University of Ottawa and IZA. E-mail: abrodeur@uottawa.ca. Cook: Wilfrid Laurier University. E-mail: ncook@wlu.ca. Neisser: University of Cologne and IZA. E-mail: neisser@wiso.uni-koeln.de. Florian Fickler, Stefan Leopold, David Winkler and Anke Witteler provided excellent research assistance. Carina Neisser acknowledges that the project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2126/1– 390838866.

The data used in applied economics has increased in both quality and quantity (Einav and Levin 2014). Access to large-scale administrative or proprietary data now provide economics researchers with opportunities to answer new and subtle questions using reliable and often representative samples (see Künn (2015) for a discussion). However, this type of data is not often shareable with the research community. While this sometimes prevents later researchers from accessing the data for their own purposes, it also prevents independent verification that results are reproducible, replicable and robust. Leveraging their position in the publication process, academic journals have begun requiring authors to submit their data at the time of publication - Christensen and Miguel (2018) document that the 'Top Five' journals in economics explicitly require data and code to be provided, with possible exemptions. The underlying motivation is that increasing the availability of data to the community enables it to reproduce or replicate the results of a prior study.[1]

While a large literature documents the extent of p-hacking and publication bias in economics and other disciplines (Adda et al. (2020); Andrews and Kasy (2019); Brodeur et al. (2016); Bruns et al. (2019); DellaVigna and Linos (2022); Doucouliagos and Stanley (2013); Elliott et al. (2022); Franco et al. (2014); Furukawa (2019); Gerber and Malhotra (2008a); Havránek (2015); Havránek and Sokolova (2020); Rosenthal (1979)), the question of whether p-hacking and publication bias depend on data-sharing policies and data availability has not received a great deal of attention. We believe this is a key research question as publication bias or p-hacking continue to cast doubt upon the credibility of published research in the eyes of policymakers (and others) with meaningful consequences. For example, if studies that identify a statistically significant effect of a given policy are more likely to be published, then this would lead to a misrepresentation of the policy's real effect (Blanco-Perez and Brodeur (2019)).

In this paper, we investigate the relationship between data-sharing policies, data (and code) availability, and data type on the presence of p-hacking and publication bias. We define p-hacking, publication bias, data availability and data-sharing policy as follows:

***p*-Hacking** refers to researcher choices being made in such a way as to manipulate or selectively report statistical significance.

Publication bias occurs when the likelihood that research is published depends upon the statistical significance of its result.

Data availability refers to whether an article's data is available either from the

[1] Availability of data is important for reproducibility of results, but also replicability (i.e., replicating prior results using the same codes but new data) and generalizability (i.e., extension of findings to other populations or settings). See Bollen et al. (2015) for definitions and a discussion of reproducibility and replicability.

publishing journal's website or from any one of the authors' homepages.[2]

Data-sharing policy refers to when a journal has implemented an editorial policy that requires authors to provide a replication package including data and code or (if granted an exemption) explicit instructions on how to obtain the data.[3]

Data type refers to a classification of an article's underlying data into four possible categories: administrative data, third-party surveys, researchers' own-collected data, and other data types, including financial data.[4]

The main hypotheses to be tested are: (1) the extent of p-hacking and publication bias in leading economics journals depends upon the existence of a data-sharing policy, and data availability; and (2) the extent of p-hacking and publication bias in leading economics journals depends upon data types.[5] We also investigate several secondary (albeit closely related) research questions such as whether p-hacking and publication bias depend on voluntary data availability on authors' homepages.

While the primary goal of data-sharing policies is not to decrease p-hacking and publication bias, but rather to enable the reproducibility and replicability of empirical results, it is plausible that these policies nonetheless decrease p-hacking through an increase in potential monitoring (we provide a more formal conceptional framework in Section 1). Data-sharing policies increase the likelihood that other interested researchers will use the uploaded replication files and thereby increase the rate of detection of (potential) p-hacking. If we assume that authors believe there will be some form of monitoring and punishment if 'caught' p-hacking, data-sharing policies might have a deterrence effect. The increased risk of being 'caught' p-hacking might change researchers' behaviour prior to publication so that p-hacking is lower in the case of strictly enforced data-sharing policies. One way to circumvent these data-sharing policies are exemptions to authors granted by the publishing journal. For instance, in case of privacy concerns attached to the release of their

[2]Due to the size of our sample, we do not check the reproducibility of the numerical results for the studies in our sample. However, we do verify whether all or only some data-sets are present.

[3]For example, the *Economic Journal*'s website states that authors must provide a replication package and that: "Authors who have requested an exemption for the publication of their datasets can either (1) grant temporary distance or physical access to the data to the journal's staff for the sole purpose of replication (the data will not be published), or (2) supply a simulated dataset or a synthetic dataset instead of the actual dataset(s) used for the analysis for replication purposes. The nature of the data used for the reproducibility checks will be indicated on the published version of the paper".

[4]Examples of administrative data include Medicare Claim Data, Tax Return Data, and Court Records. Examples of third-party survey data include the American Community Survey and the German Socio-Economic Panel. Examples of own-collected data include data from author-conducted field and lab experiments. Examples of the other data type includes those derived from Compustat and Thomson Reuters.

[5]In our context we define the extent (rather than just the presence) of p-hacking or publication bias to be related to the magnitude of our underlying test's result. For example, in the case of our caliper tests where we compare the number of test statistics above and below a threshold, the extent is related to the magnitude of the difference between these numbers.

data, researchers may not legally be allowed to share their data, making a replication of their results impossible.

To answer our research questions, we augment Brodeur et al. (2020)'s data set and collect hypothesis tests reported in journal articles employing experimental and quasi-experimental methods from 2002 to 2020 and focus on the 13 of the 25 journals who have implemented a data-sharing policy during that time. Specifically, our sample contains 38,876 test statistics published in 1,106 articles. In order to comprehensively measure data availability, we collect replication data from the publishing journals website and visit each of the authors' homepages to check voluntary data availability. For articles providing partial or no data, we collect the author-offered reasons provided by each README file. Last, we code each data set into four data types, reflective of their underlying features (e.g., administrative data is hard to access, owned by an organization, and not originally generated for research purposes).

We rely on multiple approaches to formally document the extent of p-hacking and publication bias. We start with a visual inspection of the raw distribution of z-statistics. We then follow Gerber and Malhotra (2008b) and apply caliper tests. This method focuses on discontinuities in the probability of a test statistic appearing just above or below a conventional statistical threshold. One advantage of the caliper test in comparison to other methods to detect p-hacking is that we can control for journal and year fixed effects as well as authors' and articles' characteristics. This is potentially important in our context if researchers which willingly share their data and codes have characteristics that are related to p-hacking behaviour. We also rely on a battery of p-hacking tests introduced in Elliott et al. (2022).

First, we test whether the extent of p-hacking and publication bias in leading economics journals depends upon a data-sharing policy and consequent data-availability. Using the caliper test, we do not find sufficient evidence to reject the hypothesis that sharing data and codes through a journal's webpage reduces the extent of p-hacking. Our estimates in that regard are small and statistically insignificant. We also do not find sufficient evidence to reject the hypothesis that voluntarily sharing replication material on author homepages is related to p-hacking once we control for a large set of authors and articles' characteristics. The additional battery of tests from Elliott et al. (2022) to detect p-hacking find our results robust.

Our findings are potentially driven by the fact that many articles do not share complete data and code even after the implementation of data-sharing policies. This may be partly due to authors properly complying with confidential data usage terms being exempt from providing a complete replication package. In our sample, we find

that about 60% of articles provide full data and codes in journals with a data-sharing policy. We thus employ an instrumental variable strategy, exploiting data-sharing journal policy implementation as an instrumental variable for the provision of full data and codes. While our first-stage estimates are large and statistically significant (that a data-sharing policy increases the sharing of data), our second stage again finds no evidence that data-sharing policies affect the presence of p-hacking or publication bias.

Second, we turn to testing whether the extent of p-hacking and publication bias in leading economics journals depend upon the data type. We classify data type according to colloquial categories 'administrative', 'third-party survey', 'own-collected', and 'other'. This classification reflects meaningful features of the data. For example, third-party surveys and own-collected data have the common feature that both were created for research purposes and represent relatively easier access to data, while ownership of the data differs between the two. Administrative data is hard-to-access and not originally created for research purposes, whereas own-collected data are often easier-to-access. We detail the features more fully in Section 1.

One potential disadvantage for administrative (admin) data in our setting over other data types is the relative difficulty of data access for other researchers. In our sample only 13% of administrative data are in articles which provide access to data and code for replication in comparison to 24% for third-party surveys and 55% for own-collected data. Given that a relatively large proportion of articles receiving exemptions from data-sharing policies use administrative records data, its increasing use in applied economics may raise concerns about the reproducibility of its research findings.

Nonetheless, our results suggest that the proportion of test statistics that are statistically significant (around significance thresholds) across data types is not significantly different. This result is robust to the inclusion of authors' and articles' characteristics and are consistent with additional tests of publication bias and p-hacking.

We contribute to the literature in various ways. First, we contribute to a broader literature studying the impact of journal editorial policies and the behaviour of editors and reviewers.[6] Second, we contribute to a recent literature discussing the credibility of research findings. In a recent literature review, Christensen and Miguel (2018) discuss various tools such as mandating greater data sharing and the use of

[6]See Card and DellaVigna (2020), Card et al. (2020) and Carrell et al. (2020) for recent studies documenting how reviewers evaluate papers and whether editors follow reviewers' recommendations. See Blanco-Perez and Brodeur (2020), Feige (1975) and Höffler (2017) for comments on editorial policies.

pre-analysis plans.[7] A relevant study is Brodeur et al. (2016) who document for three top economics journals that data or code availability does not mitigate p-hacking. We formalize their analysis, and extend it by analyzing this relationship for a larger number of journals, and controlling for journal fixed effects as well as authors' and articles' characteristics in our model.

Last, our results contribute to a growing literature on meta-analyses and research transparency by better informing the determinants of publication bias and p-hacking (Abadie (2020); Havránek et al. (2020); Ioannidis et al. (2017); Miguel et al. (2014); Stanley (2008); Stanley and Doucouliagos (2014); Swanson et al. (2020)).[8] Two relevant studies are Brodeur et al. (2020) and Vivalt (2019) which document differences in selective reporting by research method by providing empirical evidence that randomized control trials and regression discontinuity designs in comparison to other non-experimental methods are less p-hacked. We add to this literature by testing other potential determinants of p-hacking and publication bias in economics.

The remainder of this paper is structured as follows. In Section, 1, we provide a brief conceptual framework. Section 2 details our data collection. Section 3 investigates whether replication material availability policies decrease the extent of p-hacking and publication bias. Section 4 documents differences in the researchers who make use of admin, third-party survey and own-collected data. It also investigates whether the likelihood of providing replication material is related to data type. In Section 5, we document the impact of the "revise and resubmit" process. Section 6 concludes.

1 Conceptual Framework

In this section, we provide a brief conceptual framework providing a rationale for why data-sharing policies or certain features of data types might be expected to decrease p-hacking and publication bias.

One potential advantage of data-sharing policies is that they may change researchers' behaviour through (perceived) monitoring.[9] Authors might be less inclined to p-hack with additional monitoring following the increased likelihood that their p-hacking will be detected by the research community. This in turn could lead to a decrease in the proportion of test statistics just-rejecting the null hypothesis

[7] See Christensen et al. (2019) and McCullough et al. (2008) for a discussion of the benefits and limitations of data sharing.

[8] See Camerer et al. (2016), Chang et al. (2022), Hamermesh (2017) and Maniadis et al. (2017) among others for a discussion of replication in economics.

[9] A relevant framework is that of Becker (1968)'s crime and punishment. A p-hacker increases the presented statistical significance of their hypothesis test, but must do so under uncertainty of whether they will be detected and the severity of any repercussions.

for both submissions and published articles in journals with a data-sharing policy.

However, this mechanism assumes that authors believe there will be some form of punishment if they are caught p-hacking.[10] It may also not be possible to (easily) detect p-hacking for journals that do not strictly enforce their data-sharing policy. As of 2022, only the American Economic Association journals, the *Economic Journal* and the *Review of Economics Studies* have a dedicated data editor verifying the completeness of the replication package for the journals in our sample.

A data-sharing policy may also signal to authors that a journal's editorial board has preferences for open science practices, including positively valuing null research findings. Authors with a manuscript that does not reject the null hypothesis may then believe that their results are more likely to get published in journals with a data-sharing policy. It is thus plausible that these policies affect researchers' behaviour in the short run through redirecting the composition of submissions - at least where statistical significance is concerned. Data-sharing policies may also directly change the behaviour of editors and reviewers themselves. While it remains unclear whether reviewer preferences for statistically significant estimates have changed over time, there is evidence that adopting open science practices lead to changes in editor preferences for null results (e.g., Blanco-Perez and Brodeur (2020)).

Last, it is increasingly difficult to publish in top journals and incentives to publish in leading outlets are strong (Card and DellaVigna (2013)). These incentives may lead researchers to p-hack regardless of data-sharing policies. It thus remains unclear whether such policies could be effective at decreasing p-hacking or other questionable research practices.

In terms of data type, three features may largely determine the potential for monitoring: ease of data access, ownership of the data (by researcher or another organization), and purpose of the data (e.g. originally for research purposes or not). We thus categorize data sets in four categories informed by these features. Administrative data features more restrictive access, is owned by an organization, and was not originally generated for academic research. Third-party survey data features easier access, is owned by an organization, and was originally generated for research (whether policy or academic). Own-collection data features easy access[11] and is owned by the researchers who collected it (and generated it for their academic purposes). Last, 'other' data typically features ambiguous access, is owned by organizations, and is not originally generated for academic research.

[10] Existing reviews of published replication activities mostly document small or even minuscule replication rates (e.g., Mueller-Langer et al. (2019)).

[11] We note here that without a journal data-sharing policy own-collected data may not be easy to access for the research community. Potential p-hacking with own-collected data remains undetectable.

2 Data

In this section, we describe in detail our data collection. First, how our sample of journals and their comprising articles were chosen. Second, how test statistics were collected from articles. Third, how additional characteristics, such as data availability and data type were gathered and coded.

2.1 Journal Polices and Article Selection

Our focus is to examine the impact of journals adopting data-sharing policies, data availability, and article data-types on the presence of p-hacking and publication bias. To do this we require test statistics from many journals over a long period of time (i.e., years pre- and post-data sharing policies). We begin with the data provided by Brodeur et al. (2020) which contains 21,440 test statistics from 684 articles published in 2015 and 2018 in 25 leading economics journals.[12] The sample contains only articles using one of the following four methods: difference-in-differences (DID), instrumental variable (IV), randomized control trial (RCT) and sharp regression discontinuity design (RDD).[13] We then expand this data set as follows.

Of the 25 top journals, we identify which ones have implemented a data-sharing policy. Table 1 provides a list and their associated announcement dates.[14] In total, 16 journals in our sample had a data-sharing policy. All Top 5 economic journals have a mandatory data- and code-sharing policy.[15] The *American Economic Review* and *Econometrica* both announced their mandatory data- and code-sharing policy already in 2004. The last Top 5 journal adopting a data-sharing requirement was the *Quarterly Journal of Economics* in April 2016. Several other general interest and top field journals explicitly require data and code to be submitted at the time of article publication, including the *American Economic Journals*, *Economic Journal*, the *Journal of the European Economic Association* and the *Review of Economics and Statistics*. Last, one journal, the *Journal of Finance*, has a code-sharing policy. We code this journal as not having a data-sharing policy throughout.

We collect test statistics from articles between 2002 and 2020 (inclusive) in the exact same manner as in Brodeur et al. (2020).[16] Our additional data collection differs only with the exception that ours is a random sample (rather than the

[12] Top journals were identified using RePEc's Simple Impact Factor: https://ideas.repec.org/top/top.journals.simple10.html.

[13] Articles using matching, fuzzy RDD or Structural Equation Model are removed.

[14] We unfortunately could not obtain the announcement date for some journals where me must instead rely on the year of implementation.

[15] The Top 5 journals refer to the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*.

[16] This is done by searching the text of every article published by the selected journals in the given years for keywords related to the identification strategies (e.g. difference-in-difference* and difference in difference* for DID).

universe) of articles using the aforementioned identification methods from the 12 journals which (at some point in this almost twenty-year window) implemented a data-sharing policy.[17] Five of the journals that did not implement a data-sharing policy instead simply encourage authors to share a replication package, while three more have no specific data-sharing policies. We also note that *Economic Policy* does not publish many articles using causal identification strategies. Last, we do not include the *American Economic Journals* in this expanded data collection as all of them had a data and code policy throughout the entire time period. Nonetheless, we keep test statistics for 2015 and 2018 from these journals in our final sample. Their exclusions does not affect our main conclusions. Table 1 also provides details on the number of articles, test statistics, and years of data collection. Our final sample includes 38,876 test statistics published in 1,106 articles.

From these identified articles, we collect test statistics from within tables and only those that relate to main results. Estimates from summary statistics, balance tables, appendices, robustness checks, and placebo tests were excluded. Test statistics drawn from multiple specifications of the same hypothesis were collected. Each article was independently coded by two of the original authors to reduce concerns that only coefficients of interest were selected. All of the tests relate to two-tailed tests. In our sample, about 5%, 6% and 89% report p-values, t-statistics and coefficients with associated standard errors, respectively. We transform each of these into associated 'z-statistics'.[18]

2.2 Additional Context - Data-Sharing

We collected additional details about the availability of replications files, data accessibility and reasons for less-than-complete replication files.

To determine the availability of data and codes, we manually check every article to see if there is a replication package included on the publishing journal's webpage.[19] We also document the completeness. More specifically, we distinguish between full data and code, partial data and code, only code, and no provision at all. Note that we could not quantify the completeness of the codes for all articles in our sample, only the completeness of the package. We then also manually check

[17] We also collect test statistics from 2002 to 2020 for the *Journal of Finance*.

[18] When p-values are reported we transform them into their equivalent two-sided z-statistic values. When t-statistics are reported, we treat them as asymptotically following the normal distribution. For the most common, whereby authors report regression coefficients and standard errors, we assume the null hypothesis to be zero and so construct the associated z-statistic as the estimated regression coefficient divided by the reported standard error.

[19] Due to the manual nature of this data collection, measurement error is possible. Some articles may be considered to be 'data not provided' when in reality data is available (and vice versa, although the insightful reader may suspect that to be much less common). It is worth mentioning that we do not check for data availability or access per se, but rather if replication files are provided on journals' webpages.

every author's homepage for a replication package.[20] Appendix Table A1 provides summary statistics by data and code availability and the presence of data-sharing policy. Appendix Table A2 shows summary statistics data and code availability by journal.

Figure 1 illustrates the share of articles providing replication material for one year prior to five years after the implementation of a data-sharing policy. The year 0 is the year of the policy announcement. We restrict the sample to the balanced set of journals who ever adopt a data-sharing policy. In the year before the announcement, virtually no journal articles provided data nor codes on the journals' websites.[21] In year 1, about one-third of articles provided full data and codes, about 15% provided partial data and codes, and slightly less than 20% provided only codes. From year 3 to year 5, about 20% of articles still did not provide codes nor data. Approximately 60% of articles provided full data and codes.

To determine data accessibility, we manually check each replication package and corresponding README file. Graphically presented in Appendix Figure A1, we see that from between year 3 to 5 about 60% of replication packages provide direct access to the data. An additional 20% of authors offer some forms of help or guidance on how to access the data. The remaining 20% of articles do not provide any access nor guidance on how to obtain the data for replication.

To determine the reasons for less-than-complete replication data, we examine the README file and check whether a reason was provided for incomplete or missing data.[22] We categorized exemption reasons into the following categories: (1) need approval (i.e., authors encourage interested researchers in writing an application and provide contact details), (2) need approval and paying a fee (i.e., authors explicitly mention a necessary payment), (3) confidential data (i.e., authors use confidentiality reasons/proprietary data as an excuse and provide no help or contact details in accessing the data), and (4) non-distributable data (i.e., authors are not allowed to share the data but they provide a link, where the data can be accessed). Multiple exemption reasons may be provided for a given study. For example, a study may use two datasets, with different reasons for not sharing each dataset. For articles published in journals with a data-sharing policy that do not fully provide their

[20] In total, we accessed 98,796 homepages (including duplicate access for repeated authors). In our sample, about 18% of articles had replication material on at least one of the authors' homepage. This figure goes down to 12% for articles published in journals without a data-sharing policy. Appendix Figure A2 illustrates that the share of authors sharing data and codes on their personal homepages does not meaningfully change after the implementation of a data-sharing policy.

[21] One article published in the *Quarterly Journal of Economics* released full data and codes on the journal's website prior to the implementation of the data-sharing policy.

[22] We code data accessibility as reported in the README files and check for data access at the time of publication. Of note, this may lead to measurement error in instances where the data become public-use over time (e.g., court records obtained via multiple FOIA requests).

data, only 53% of authors provide exemption reasons for not sharing (some or all of) their data. For those that do provide a reason, the majority (76%) mention "need approval". Only 7% mention "need approval and paying a fee", while about 17% and 20% mention confidential data and not distributable, respectively. We note that exemption reasons are often interchangeable and that measurement error is likely. Readers should thus be careful when interpreting these findings.

2.3 Additional Contextual Data

We follow Brodeur et al. (2020) and collect additional contextual data. For each article, we record: the journal, year of publication and the number of authors. For each author, we record: gender, current institution, PhD granting institution, year of graduation, and whether the author was an editor of an economics journal at the time of publication.[23] These included article and author characteristics are the same as in Brodeur et al. (2020). Table 2 provides descriptive statistics for article and author characteristics by data availability. The unit of observation is a test statistic. We see that about 34% of tests in our sample come from articles published in the Top 5 journals in economics. Approximately 11% of tests come from solo-authored articles. The average years of experience of authors (years since PhD completion) in our sample is 10.9 and about 23% of tests are in articles written by authors affiliated with a top institution.[24] We find that articles providing full data and codes on the journals' website are more likely to be published in a Top 5 journal. Similarly, there appears to be a positive relationship between years of experience (and affiliation ranking) and providing full data and codes. In contrast, data-sharing practices do not seem to be related to gender and number of authors.

2.4 Additional Context - Data-Type

For each article, we collect information on data sets used. More specifically, we collect information on the method of data collection and the name of the data set.

When classifying our data-types we used the following guidelines, characterized by certain features of the data. Administrative data is originally collected for a purpose other than for academic research. Third-party survey data is collected by organizations (including governments) for later research use. Own-collected data

[23] We record gender using self-reported information in CVs and authors' homepage biographies and head-shots. Other authors' characteristics are collected from authors' homepage biographies as well.

[24] We follow Brodeur et al. (2020) and code as "top" institutions the following institutions/departments: Barcelona GSE, Boston University, Brown, Chicago, Columbia, Dartmouth, Harvard, MIT, Northwestern, NYU, Princeton, PSE, TSE, UC Berkeley, UCL, UCSD, UPenn, Stanford, and Yale. The choice of institutions was based on RePec's ranking of top institutions (https://ideas.repec.org/top/top.econdept.html).

is collected by researchers (or individuals under direct researcher supervision). If none of these labels apply, we classify the dataset as "other". We note that this manual classification may include misclassification. While we believe some misclassifications are more likely than others, we believe that misclassification of any type into the "other" category is most likely, and so remain reserved when interpreting any conclusion based on that data-type.

Administrative (or register) data are generally collected by government agencies used for administrative purposes. Typical examples are social security or vital records. Compared to administrative data, *third-party survey* differ in terms of their purposes. Third-party surveys are conducted to answer specific questions, while often targeting only subgroups of individuals. For example, a candidate survey conducted during an election. Two prominent examples are the Current Population Survey (CPS) and the General Social Survey (GSS). These data are gathered by a third party and not by later academic researchers themselves.[25] *Own-collected data by researchers* on the other hand, describe data sets that are manually collected by researchers or research assistants. Such data might be an own-implemented survey or field experiment. We coded all remaining data sets as *other*. This involves data collected from financial data streams such as Bloomberg or Compustat but also statistical data like GDP or unemployment figures that are publicly available and provided by organizations such as the OECD or World Bank.

For our data type analyses, we restrict the sample to articles relying solely on one type (we refer to the moniker 'pure sample'). Appendix Figure A3 illustrates the evolution of data type use over the entire time period. We document a large increase in the use of admin and own-collected data, and a sharp decrease on the use of third-party surveys. Table 3 provides summary statistics for the type of data used in our sample and Appendix Table A3 further describes data type by data and code availability.[26] In total we identified 20,701 observations and 597 articles that rely on solely one data type. The largest share of tests is collected by the researchers themselves (37.5%), while approximately 27.7% employ admin, 19.0% third-party survey and 15.9% rely on other data, respectively. We also see that about 40% of articles using admin and own-collected data are published in Top 5 journals against 20% for third-party survey data. Articles using own-collected data are more likely to be solo-authored and have more years of experience on average.

If we restrict the sample to articles which rely solely on one data type (i.e., pure sample) and provide full data and code, we observe some differences that confirm

[25] See Kapteyn and Ypma (2007) for a discussion of issues and problems with third-party survey and admin data. The authors point out, for instance, that surveys are more costly and subject to non-response issues, while admin data may suffer from mismatching due to imperfect linkage information from different sources.

[26] Appendix Table A4 also provides an overview of data type by journal.

some of our anecdotal expectations. Tests from research that use administrative data are the least likely to provide full data and code at 5.4%. Research published using data that researchers collected on their own is the largest share of those that provide full data and code at 66.2%. Third party surveys and other data make up 16.7% and 11.7% respectively. These differences are not driven by outliers – if we consider the sample of *articles* that provide full data and code, we observe the same pattern. Across those articles that provide full data and code, 12.3% use admin, while 54.6% rely on own-collected data by researchers. This result is consistent with two possibilities; (1) a large share of admin journal articles receiving exemptions from data-sharing policies at top journals and/or (2) through composition effect in which admin data papers are more likely to be published in journals that do not have data-sharing policy.

3 Data-Sharing Policy, P-Hacking and Publication Bias

We first visually investigate the distribution of z-statistics for journals with and without data-sharing policies. In a second step, we apply more formal approaches to detect p-hacking and publication bias.

It is useful to clarify what the distribution of z-statistics 'should' look like in the absence of p-hacking and publication bias. Brodeur et al. (2016, 2020) noted that in a region where the incentives for a researcher to misrepresent statistical significance are small, the observed distribution of test statistics closely resembles a student's t-distribution. More recently, Elliott et al. (2022) derive theory that for any underlying distribution of true effects a p-curve (a mechanical equivalent to the z-statistic distribution) should be non-increasing and continuous without the presence of p-hacking or publication bias.

In our setting, if publication bias is present (that is the publication process at some point prefers higher z-statistics to lower ones, as in Brodeur et al. (2016)) we expect a rightward shift of mass in the z-statistic distribution that maintains the monotonically decreasing aspect of the z-statistic distribution. On the other hand, p-hacking could result in too many z-statistics that are just above a threshold (Simonsohn et al. (2014)) or too few z-statistics below a threshold (Brodeur et al. (2016)). If either (or their combined) effect is large enough the monotonically decreasing aspect of the z-statistic distribution could be violated, and most likely around a statistical significance threshold where researchers perceive a benefit to just crossing. Under p-hacking, we expect to see either a valley, a peak, or both around a threshold.

3.1 Distribution of z-Statistics

In what follows, we illustrate the raw distribution of z-statistics for several subsamples for $z \in [0, 10]$. Similar to Brodeur et al. (2020), we create z-curves by estimating kernel densities. A kernel smooths the distribution, softening both valleys and peaks. Reference lines are provided at conventional two-tailed significance levels.

Data-Sharing Policies. Figure 2 illustrates the distributions of test statistics for three subsamples. The first panel restricts the sample to articles published in journals with no data-sharing policy. The second panel restricts the sample to articles published in journals with a data-sharing policy. The third panel restricts the sample to articles published in journals only encouraging data-sharing. Later, we treat those observations with an encouraging data policy as having no journal policy. We include all years in our sample. Appendix Figure A4 plots the three z-curves into a single panel for ease of comparison.[27] To the eye, the distributions are remarkably similar[28] exhibiting a non-monotonic peak at about 1.96 after a valley between 1.5 and 1.65. More precisely, the z-curves for articles in journals with and without a data-sharing policy are right on top of each other for the entire distribution. This is a first piece of evidence that data-sharing policies have little impact on p-hacking and publication bias.

Data and Code Availability. We further investigate this relationship by restricting the sample to journals that implemented a data-sharing policy. Figure 3 plots the distribution of z-statistics from the articles providing full data and code, partial data and code, only code, and no replication material separately. The four panels reveal qualitatively similar patterns[29] each with an apparent peak around 1.96. The peak appears to be a bit sharper for articles providing full data and codes. This is a second piece of evidence that data-sharing policies do not meaningfully decrease p-hacking.[30]

[27] In the appendix we provide figures plotting multiple z-curves into a single panel for ease of comparison. Referring to 'Data-Sharing Policies', see Appendix Figure A4. Referring to 'Availability of Replication' Material, see Appendix Figure A5. Referring to 'Timing of the Implementation of Data-Sharing Policies' see Appendix Figures A6, A7 and A8.

[28] Despite a two-sided Kolmogorov–Smirnov test comparing no policy to a policy returning a p-value of 0.025, no policy to encouragement of 0.0298, and policy to encouragement of 0.014.

[29] A Kolmogorov–Smirnov test comparing full data and code to partial data and code returns a p-value of 0.110, partial data and code to only code 0.074, and only code to no provision 0.035.

[30] We explore heterogeneity effects across different subsamples in Appendix Figures A9 – A12. These figures illustrate decompositions by journal ranking (Top 5 and non-Top 5), number of authors, institutional rank and PhD institutional rank. Test statistics in Top 5 journals seem to behave the same as non-top 5 in all replication availability material settings, although partial data and code does look flatter for Top 5. This pattern repeats with multi versus solo authorship with solo-authored partial data flatter than multi-authored test statistics. A high ranking institution seems to differ from others when only code is provided, this is less pronounced when we consider a high ranking PhD granting institution.

Timing of the Implementation of Data-Sharing Policies. Last, we visually investigate the timing of the implementation of data-sharing policies. We start by looking at whether the distribution of z-statistics changes from before to after the implementation of a data-sharing policy. Figure 4 plots the distribution of z-statistics for the 5 years prior to the announcement of the policy (left panel) and the distribution of z-statistics for the 5 years after the announcement (right panel), respectively. Visually, there does not seem to be any discernible change from before to after the announcement and implementation of the data-sharing policy.[31] To determine if a data-sharing policy becomes more effective at reducing bunching near statistical significance thresholds over time, we decompose the post-policy years into two categories: first two years vs third to fifth year after policy implementation. Again the two z-curves are mostly on top of each other for most of the distribution, although there seems to be slightly less bunching around 1.96 for third to fifth year after policy (see Appendix Figures A7 and A8).

Voluntary Provision by Authors. We also explore whether the distribution of z-statistics differ depending on the availability of any replication material at all, including authors' homepages. In Appendix Figure A16, we plot the distribution of z-statistics for articles for which replication material is available on at least one of the authors' homepages. As a comparator, we plot the distribution of z-statistics for articles in which no material was available on the journal's webpage nor any of the authors' homepages. There is little discernible differences between the two panels. Overall, our visual inspection seems to suggest that neither voluntary data-sharing nor data-sharing policy have an impact on the distribution of z-statistics.

Reasons for Data Exemptions. We last investigate whether the distribution of test statistics is similar for authors providing a reason for not sharing data in comparison to those not providing a reason. For this exercise, we restrict the sample to articles published in journals with a data-sharing policy and have partial or no data. The distributions are illustrated in Figure 5 and Appendix Figure A17. These figures show that the two kernel density lines are on top of each other for most of the distribution, with slightly more mass near 0 for articles without a reason. We interpret those findings as suggestive evidence that providing a reason or not is not related to p-hacking or publication bias.

[31] One potential issue is the overrepresentation of round values (e.g., coefficient of 0.02 and standard error of 0.01). We follow Brodeur et al. (2016) and deal with this potential issue by randomly redrawing a number in the interval of potentially true numbers around each collected value using a uniform distribution. This de-rounding method has no impact on our conclusions. See Appendix Figures A13-A15.

3.2 Methods and Tests for Detecting P-Hacking

In this subsection, we formally test whether data-sharing policies decrease the extent of p-hacking and publication bias. We first describe and present results based on the caliper test, which consists of comparing test statistics close to significance thresholds (Gerber et al. (2008)). We then conduct additional tests designed to detect p-hacking introduced by Elliott et al. (2022) before turning to testing whether data-sharing policies decrease publication bias using the methodology developed by Andrews and Kasy (2019). Of note, the caliper test method and the battery of tests from Elliott et al. (2022) jointly identify p-hacking and publication bias in the additional presence of publication bias.[32]

3.2.1 Caliper Test The caliper test compares the number of test statistics in a narrow equal-sized range above and below a statistical significance threshold. 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. Under the null hypothesis that the number of test statistics should be equal above and below a statistical threshold[33] if there is a sufficiently large difference in the numbers the probability the difference is due to chance is small and we may reject the null hypothesis.

We focus throughout on the 5%, but provide similar analyses for the 1% and 10% thresholds.

Relationship between data-sharing and the *p*-hacking We estimate the following equation to estimate the relationship between data-sharing and the likelihood to report a statistically significant result:

$$Pr(Significant_{iajt} = 1) = \Phi(\alpha + \beta_j + \gamma_t + \lambda DataProvidedJournal_{ajt} + \mu DataProvidedAuthor_{ajt} + X'_{iajt}\delta) \quad (1)$$

where $Significant_{iajt}$ is a dummy variable for whether test i in article a in journal j in year t is statistically significant at the 10%, 5% or 1%-level. We rely on probit models throughout and present the average marginal effects and associated standard errors clustered at the journal article-level. The variables of interest are $Data\ Sharing\ Journal_{ajt}$, which represents a dummy variable for whether the authors shared full data and codes on the journal's website, and $Data\ Sharing\ Author_{ajt}$,

[32] Consistent with the current literature, any method we employ in this paper is a joint test for p-hacking and publication bias when in the presence of both. In our analysis, we consider a broad set of articles, across a large set of journals, and across a significant period of time which in our opinion is more likely to contain both than the absence of either.

[33] Equally, that there is nothing influencing whether a test statistic is just below or just above a threshold.

which represents a dummy variable for whether one of the authors provided a replication package on their homepage. These variables equal zero if the authors provide partial data, only code, or no replication package at all on the journal's webpage and author homepages, respectively. The main advantage of using caliper test instead of a graphical examination of the distribution of z-statistics is that we can control for authors' and articles' characteristics. The vector X'_{iajt} denotes a set of covariates identical to the ones used by Brodeur et al. (2020), which includes dummy variables for how results are reported (i.e., p-values, standard errors or t statistics), 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 (and its square), average PhD (granting) institutional rank, average (current) 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 supplement the set of covariates by adding identification strategy dummies (e.g., instrumental variable) as well as both year and journal fixed effects.

Estimates of equation (1), with the 5% significance thresholds as the dependent variable, are reported in Table 4. In columns 1 and 2, we restrict the sample to $z \in [1.46, 2.46]$ for the 5% statistical significance threshold. We present estimates for successively smaller bandwidths in columns 3–6 (i.e., $z \in [1.61, 2.31]$ and $z \in [1.76, 2.16]$).[34] In column 2, 4 and 6, we add authors' and articles' covariates to account for any characteristics that might also be related to p-hacking behaviour.

In Table 4, the estimated coefficients for data-sharing on the journal's website, λ, are small, positive and statistically insignificant in all columns. This is not consistent with the hypothesis that sharing full replication material reduces the extent of p-hacking.[35] Appendix Tables A5 and A6 find the same for the 10% and 1% levels.[36]

The estimated coefficients for sharing replication material on at least one of the authors' homepage, μ, are small, negative and statistically significant solely

[34] We note here that the assumed equality of mass above and below a threshold holds in the limit (as the bandwidth approaches zero). We have chosen our bandwidth both from convention (Gerber and Malhotra (2008a), Brodeur et al. (2020)) and considerations towards having sufficient test statistics to operationalize the test.

[35] An additional assumption being made is subtle but meaningful. When examining the effect of an author or article characteristic in the light of the caliper test, we are assuming that the characteristic in question does not meaningfully change the probability of a test statistic being *just* over or below a statistical threshold. For example, test statistics which have their data shared on a journal's website or of different data types may have an underlying distribution (absent p-hacking and publication bias) that differs from the expected equality (or perhaps may even increase) in the small interval surrounding a statistical threshold. While we cannot rigorously rule out such discontinuities from occurring 'naturally', in terms of the variables of interest we consider (provision of replication materials, underlying data type) we do not have a compelling reason for a strong discontinuity to exist other than through the p-hacking or publication bias mechanisms - but we cannot definitely rule out the possibility either.

[36] We present estimates for the main control variables in Appendix Table A7.

in columns 1 and 3. The estimates are smaller in magnitude and statistically insignificant in columns 2 and 4 when we add the authors' and articles' covariates, suggesting that voluntarily sharing a replication package may not be random and may be related to observable authors' and articles' characteristics. Notably, the estimates are small and insignificant in columns 5 and 6 where we rely on a smaller bandwidth. For the 10% threshold, the same pattern occurs with only marginal significance in the largest bandwidth, which disappears following the introduction of article and author controls and smaller bandwidths. For the 1% threshold, μ is never statistically significant, and is also inconsistently signed.

We take a moment here to discuss whether these results are meaningfully well-estimated null effects or if our results are statistically insignificant simply due to noise in the data. If we consider a 95% confidence interval, we can rule out that data and code sharing at the journal's webpage does not decrease p-hacking by more than 7.8% and that voluntary data and code sharing at any author's homepage does not decrease p-hacking by more than 2.7%.

Data-Sharing Policy and *p*-hacking We now turn to estimating the relationship between a data-sharing *policy* and p-hacking behaviour. The equation we are estimating is similar to equation (1), but we replace the variable $Data\ Sharing_{ajt}$ with the dummy variable $Data\ Sharing\ Policy_{jt}$ indicating whether journal j had announced or implemented a data-sharing policy in year t. This analysis can be thought of as a 'reduced form' in a more conventional instrumental variables estimation we conduct momentarily. These results are presented in the third row of Table 5. For this analysis, we restrict the sample to journals with a data availability policy and to five years pre- and post-policy. No estimate is statistically significant and are inconsistently signed depending upon the bandwidth chosen.

Instrumental Variables Approach We now rely on an instrumental variable strategy to investigate if data-sharing reduces p-hacking and publication bias. Specifically, we require an instrument that is correlated with the likelihood to share data and affects the likelihood to report statistically significant results around significance thresholds only through data-sharing (i.e., exclusion restriction). We thus use the data-sharing policies as an instrument for data-sharing. There are a number of reasons why the exclusion restriction may not be satisfied. For instance, data availability policies might impact the types of submissions received by journals who implement them. For this reason, we also report IV estimates after controlling for our full set of authors' and articles' characteristics. However, the results should be treated with caution when interpreting as causal.

We estimate the following equations:

$$\begin{cases} Pr(DataSharing_{iajt} = 1) = \Phi(\rho + \beta_j + \gamma_t + \phi \cdot DataSharingPolicy_{ajt} + X'_{iajt}\psi) \\ Pr(Significant_{iajt} = 1) = \Phi(\alpha + \beta_j + \gamma_t + \lambda \cdot \widehat{DataSharing}_{ajt} + X'_{iajt}\delta) \end{cases} \tag{2}$$

The first-stage is reported in the first panel of Table 5. (See Appendix Tables A8 and A9 for the 10% and 1% significance thresholds.) Unsurprisingly, we find that data-sharing policies meaningfully increase the likelihood of sharing full data and codes. The point estimates are all statistically significant at conventional levels regardless of control inclusion and bandwidth selection.[37]

The second-stage estimates are reported in the middle panel of Table 5. Consistent with the reduced form and the results from Table 4, the second-stage estimates are all statistically insignificant; data-sharing (even when instrumented using data-sharing policies) do not change the share of significant estimates published. The sign is positive in two columns and the 'expected' negative in four columns. This result is robust to different bandwidths and the inclusion of control variables, significance thresholds.[38]

3.3 Further Tests for P-Hacking

Elliott et al. (2022) formalized the expectations of the test statistic distribution under the hypothesis of no p-hacking. They provide testable conditions of the p-curve (a histogram of p-values) in comparison to our primary analysis of z-statistics. We are motivated by a desire for reader ease and so discuss our analysis in terms of p-curves in this section.

We apply the code provided by Elliott et al. (2022) to our full dataset as well as sub-samples of interest; each application provides results from six different tests of p-hacking: Binomial, Fisher's, Discontinuity, CS1, CS2B, and LCM.[39]

For this section, we apply the methodology to four divisions of our data and present them in Table 6. First, for all journals and years we compare p-hacking and publication bias by whether full data and code were provided on the journal website. Second, we restrict the sample to only those journals that adopted a data and code policy and examine up to 5 years *before* its implementation by whether

[37] We estimate OLS models in Appendix Table A10 and report Cragg-Donald Walt F-statistics. The F-Statistic ranges from 105 to 144 in our baseline model.

[38] We show that our results for the 5% significance threshold are robust to the use of article weights and de-rounding in Appendix Tables A11-A14.

[39] For the most salient statistical significance threshold of $z = 1.96$, we have no need to modify the replication code provided by Elliott et al. (2022). We also conduct the same battery of tests for the 10% and 1% thresholds. Notable, only the Binomial and Discontinuity tests require re-application, as the other tests examine the broad histogram rather than a single point. The results of those tests can be found in Appendix Table A15.

full data and code were provided. Third, we restrict the sample in the same way but examine up to 5 years *after* the data and code policy was implemented. Fourth, we examine for those journals that do implement a policy both before and after their implementation. For context, Elliott et al. (2022) considered any p-value less than 0.10 to be evidence of p-hacking in the original article.

The binomial test, as originally operationalized, compares the mass $p \in (0.045, 0.050]$ (those test statistics that are just statistically significant) to the mass in $p \in [0.040, 0.045]$ (those that are just slightly more statistically significant). Under the null of no p-hacking (and no publication bias) the histogram of p-values should be non-increasing, in other words the mass of $p \in [0.040, 0.045]$ should be greater than the mass of $p \in (0.045, 0.050]$. Visually, this would correspond to a histogram bar just to the left of the threshold that is too-tall compared to its left-neighbour.[40] For the entire sample, we can reject the null hypothesis of no p-hacking and no publication bias with great confidence ($p < 0.000$). In comparing each of the four divisions, there seems to be little difference in the results of the binomial test.

The discontinuity test is an application of Cattaneo et al. (2020) using data-driven bandwidth selection (Cattaneo et al. (2021)). When examining the full sample, those articles published with replication materials exhibit a discontinuity ($p = 0.002$) whereas those without are not ($p = 0.760$). This result highlights the benefits to restricting our sample - as soon as we examine instead the post-policy period, the replication materials articles have 'marginally weaker' evidence of a discontinuity ($p = 0.082$) as compared to Other ($p = 0.001$).

The CS1 (non-increasingness) and CS2B (bounds on the p-curve and its first two derivatives) tests are both histogram-based tests introduced by Elliott et al. (2022) and are more powerful than the more commonly used binomial and Fisher's - particularly in situations where p-hacking need not violate the non-increasingness of the p-curve.[41] Perhaps due to this increased power, we receive a statistically significant result regardless of division.

The LCM test, which attempts to reject the null that the CDF of the p-curve is concave (a direct consequence from the property that the p-curve itself is non-increasing), does not find large differences between whether or not replication materials are provided, regardless of division.

In summary, while there may be some suggestive evidence that comes from one test (the Binomial) that providing data and code *post* policy reduces p-hacking, the tests we apply from Elliott et al. (2022) do not detect a consistent or definitive difference between the statistical behaviour of articles which provide data and code

[40] Notably our calipers compare the just-above to just-below significance masses of test statistics, making the addition of the Elliott et al. (2022) analysis valuable along an additional dimension.

[41] We have omitted an analysis of Fisher's as it returns $p = 1.000$ regardless of application.

and those who do not.

3.4 Publication Bias

Andrews and Kasy (2019) provide a framework and methodology aimed at identifying and correcting for publication bias. As a reminder, publication bias specifically refers to the statistical significance of a result affecting the probability of that result being published. While their key findings are that corrected-meta analyses and replication studies offer similar results (in the topics where both are available), their developed methodology allows us to examine relative publication probabilities. That is, we can compare whether the statistical significance of a result affects whether it was published. This allows us to assess the extent of publication bias, i.e., the magnitude of the relative publication probability.

We apply the code provided by Andrews and Kasy (2019) in an unchanged form to four divisions of our sample: First, for all journals and years whether full data and code replication materials were provided on the journal website against any Other. Second, we restrict the sample to only those journals *who would eventually adopt a data and code policy* and examine up to 5 years before its implementation by whether full data and code were provided. Third, we restrict the sample in the same way but examine up to 5 years *after* the data and code policy was implemented. Fourth, we examine for those journals that do implement a policy both before and after their implementation. This method involves applying a step function at statistical significance thresholds, we choose to model the 10%, 5%, and 1% thresholds jointly. We assume that the underlying distribution should follow a generalized t-distribution.

Table 7 presents the estimates for the relative publication probability of a test statistic as compared to a baseline or reference test statistic that falls in the interval $z > 2.58$. The first three columns of the table present the parameters of the underlying distribution fit by the model. The first column fits the 'mean effect', which, in the original context of meta-analysis would be a literature's estimate of the underlying effect (for example the effect of minimum wage on employment from a number of independent studies). We do not provide an interpretation of this variable as our estimates come from studies estimating different effects - it is simply the location parameter - however we include it for later replicators. The second column provides the scale parameter and the third column presents the degrees of freedom. The remaining columns present our main estimates; for the full sample a statistically insignificant result is just over a third (35.7%) as likely to be published as one that is statistically significant at the 1% level. A significant result with one star is around 79.1% as likely. Interestingly (and not an anomaly to our study)

those test statistics that are statistically significant at the 5% level are more likely to be published than one at the 1% level - indicative of publication bias.

Whether we examine the full sample, up to 5-years before a journal data and code policy is implemented, up to 5 years after a journal data and code policy is implemented, or 5 years pre-and-post-policy, there seems to be little difference between those articles (tests) that provide data and codes and others. Once again, we have suggestive but not definitive evidence of a reduction in publication bias following the implementation of a data and code policy - noting the *change* of relative publication probabilities from around 1.5 to around 1.1 of two-star as compared to three-star results following the implementation of the policy.[42] To sum up, we find limited evidence that data-sharing policies significantly reduce the extent of publication bias in economics using the method and code provided by Andrews and Kasy (2019).

4 Data Type, P-Hacking and Publication Bias

Our analysis of data-sharing policies uncovered no reduction in p-hacking and publication bias. In this section, we first investigate whether data-sharing practices vary across types of data. As a preview, we find that articles using administrative data are less likely to share full data and codes than own-collected data. We then investigate whether different types of data may suffer from more p-hacking. A lack of differences in p-hacking behaviour across data types could provide a plausible explanation for our lack of significant results on the effectiveness of data-sharing policies in reducing p-hacking.

Throughout this section, we restrict the sample to articles relying solely on one data type (a 'pure' sample). This decision was made to avoid issues related to studies with multiple data types (e.g., dependent variable uses admin data while independent variables use survey data).[43]

[42] While one-star results are also less likely to be published compared to three-star results following a policy, this move is also accompanied by a reduction in publication probability of statistically insignificant (no-star) results, suggesting that the policy may have only increased publication probabilities of very significant results. This is further noted by the relative stability of the relative publication probability ratios between one-and two-star results before and after a policy.

[43] We provide evidence that the omission of journal articles using multiple types of data has no effect on our conclusions. In Appendix Figure A18, we plot two z-curves into a single panel. The first z-curve restricts the sample to journal articles that rely on one type of data, while the second curve does not impose this restriction and rely on the full sample. The distribution for both samples is extremely similar, and both exhibit a peak between 1.65 and 2.5. See Appendix Figure A19 for the de-rounded distributions.

4.1 Data Type and Replication Material Availability

We first test the relationship between types of data and replication material availability. We estimate the following equation:

$$Pr(DataSharingJournal_{iaft} = 1) = \Phi(\alpha + X'_{iat}\delta + \gamma Survey_{iat} + \lambda OwnCollected_{iat} + \theta Other_{iat}) \quad (3)$$

where $DataSharingJournal_{iaft}$ is a dummy variable for whether the authors shared full data and codes on the journal's website for test i in journal article a in field f in year t. We rely on probit models and present standard errors clustered at the journal article-level. The variables of interest are $Survey_{ia}$, $Own\ Collected_{ia}$ and $Other_{ia}$, which represent dummy variables for different data types. Administrative data is the reference category, which is omitted.

We include in our model the term X_{iat}, which is a vector of articles' and authors' characteristics. We include our usual set of covariates, which includes dummy variables for methods (e.g., instrumental variable). The inclusion of method dummies is very important given that own-collected papers are often randomized controlled trials. A key remaining concern is that unobservable characteristics could be related to data types and p-hacking behaviour. Readers should thus view our analysis here as exploratory and particularly non-causal.

The results are presented in Table 8. In columns 1–3, the dependent variable is whether full data and codes are provided on journals' webpages. For columns 4–6, the dependent variable is a dummy variable indicating that at least the codes for replication are provided on journals' webpages. In columns 1 and 4, we include only our variables of interest for data type. Columns 2 and 5 add to the model articles' and authors' characteristics. In columns 3 and 6, we also include field fixed effects. More precisely, we include dummy variables for the following fields: general interest, finance, macroeconomics, development, experimental, public and urban economics.

We find that third-party surveys and own-collected data are significantly more likely to provide data and codes than the reference category of admin data.[44] Our estimates in column 3 suggest that third-party survey and own-collected data are about 25 and 40 percentage points more likely to provide full data and codes than admin data, respectively. The estimates are statistically significant at conventional levels. For columns 4–6, we also find that studies relying on own-collected data are more likely to provide at least codes for replication than studies using admin data.

[44] A related question is what predicts the provision of data and codes in the presence of a data-sharing policy? Appendix Table A16 replicates Table 8 but we restrict the sample to journals with a data-sharing policy. Our conclusions are similar with articles using admin data being less likely to provide full data and codes.

In contrast, we do not find evidence that studies relying on third-party survey data are significantly more likely to provide at least codes for replication compared to admin data – estimates are small and statistically insignificant in all models.

The estimates for most of the control variables are not statistically significant. One exception is the Top 5. The estimates are statistically significantly and suggest that articles published in these outlets are more likely to provide full data and codes. This result is consistent with the fact that Top 5 journals all had a mandated data sharing by 2016 (Christensen and Miguel (2018)).

We also investigate the reasons for not providing full data by data type. We restrict the sample to articles published in journals that have a data-sharing policy and did not provide full data. For admin and third-party survey data, we find that about 74% and 58% of articles do provide a reason for not sharing all their data, respectively. This is against only 12% and 35% for own-collected and other data, respectively. The most popular reason given for all data types is need approval (i.e., authors encourage interested researchers in writing an application and provide contact details). We take from this the following conclusion; many authors that use admin data are aware they cannot release their data, but aim to alleviate this lack of monitoring by offering to help interested researchers apply for the data.

4.2 Distribution of z-Statistics Across Types of Data

Figure 6 displays the raw distribution of z-statistics for each of the four data types.[45] The shapes are striking with all featuring a hump around the 5% significance threshold. For third-party survey data, the distribution exhibits a local minimum around 1.5 and a maximum around 1.96. Approximately 54.7%, 46.2% and 30.6% of test statistics are significant at the 10, 5 and 1 percent levels, respectively. The distribution of z-statistics for admin data also exhibits a local maximum around 1.96. About 56.9%, 50.3% and 37.2% of test statistics are significant at the 10, 5 and 1 percent levels, respectively. In contrast, own-collected data displays an almost monotonically falling curve with a much smaller local maximum around 1.96. 48.2%, 41.0% and 27.2% of own-collected test statistics are respectively significant at the 10, 5 and 1 percent levels.[46]

[45] In Appendix Figure A20, we plot the z-curves into a single panel. In Appendix Figure A21, we deal with the overrepresentation of round values. The only noticeable change is less mass at exactly $z = 2$ for the own-collected distribution.

[46] We further investigate these patterns for different subsamples in Appendix Figures A22-A25. These figures illustrate decompositions by data type by journal ranking (Top 5 and non-Top 5), number of authors, institutional rank and PhD institutional rank. Of note, we find that the spike around 1.96 is more pronounced for journal articles with no authors who graduated from a top university for admin data users. In contrast, current institutional rank does not appear to be related to the spike around the 5% threshold for all data types. The shape of the distributions is quite similar for solo- and multi-authored articles, with the exception of own-collected data where the spike is particularly striking for solo-authored articles.

Last, the distribution of z-statistics in journal articles categorized as using data in the Other category has a global and local maximum around 1.96 and seems to suffer the most from p-hacking and/or publication bias. 72%, 65% and 48% of Other test statistics are significant at the 10, 5 and 1 percent levels, respectively. Appendix Figure A26 splits the Other data type into two categories: financial data and non-financial data. Among the test statistics in Other, about half are in articles relying on financial data. This split into financial and non-financial data illustrate that the distribution of z-statistics for both these subgroups is remarkably similar (Kolmogorov–Smirnov test, p=0.139).

4.3 Methods and Tests for Detecting P-Hacking

We show caliper estimates for the 5% significance threshold in Table 9. (See Appendix Tables A20 and A21 for our estimates for the 10% and 1% significance levels.) The table has the same structure as Table 4. One key difference is that our variables of interest are dummy variables for data types instead of dummy variables for the provision of full data and codes. The coefficients presented are increases in the probability of statistical significance relative to the baseline category (admin). Our sample is smaller as we now restrict the sample to articles using only one type of data, as noted earlier.

In columns 1 and 2, we restrict the sample to $z \in [1.46, 2.46]$ for the 5% statistical significance. We present estimates for smaller bandwidths in columns 3–6 (i.e., $z \in [1.61, 2.31]$ and $z \in [1.76, 2.16]$). We include year and journal fixed effects in all columns and our usual set of control variables in even columns. We also control for the identification method for two reasons. First, Brodeur et al. (2020) provide suggestive evidence that studies relying on instrumental variables as a method are more p-hacked than studies using RCTs or RDD. Second, in our sample 85% of own-collected data rely on RCTs.[47]

Overall, we find no evidence that specific data types suffer differently from p-hacking or publication bias. The point estimates for hand collected, third-party survey and other data are very small in magnitude.[48]

[47] Appendix Figures A27 - A30 illustrate the distribution of z-statistics by method of data collection for difference-in-differences, instrumental variables, randomized control trials and regression discontinuity design, respectively. Similarly, Appendix Figures A31 - A34 illustrate the distribution of z-statistics by availability of replication material for difference-in-differences, instrumental variables, randomized control trials and regression discontinuity design, respectively. We include these figures for the interested reader but caution that the small sample sizes make drawing conclusions tenuous at best.

[48] We also apply a suite of tests from Elliott et al. (2022). The results are presented in Appendix Table A17. The binomial test finds evidence of either publication bias or p-hacking for all data types, while the discontinuity test does not find evidence of either for all data types. Both CS1 and CS2B (tests of the non-increasingness of the p-curve and considered more powerful than the others by Elliott et al. (2022)) detect p-hacking or publication bias for administrative, own-collected, and

4.4 Publication Bias

In this subsection we apply the model of Andrews and Kasy (2019) in order to examine publication bias using relative publication probabilities.[49] In Appendix Table A22, we present the results of applying the model separately to each of our four data type sub-samples. That is, we compare the relative publication probabilities *within* data type. The structure of the table is the same as Table 7. We find that a statistically insignificant test statistic derived from a study using administrative data (where $0 < t < 1.645$) is around 43.5% as likely to be published as one from the omitted category of very statistically significant test statistics (where $t > 2.576$). In magnitude this is roughly similar for own-collected data. More severely, third-party survey and other data sources which provide statistically insignificant test statistics are 28.8% and 20.2% as likely to be published - around a quarter to a fifth as likely - evidence of a relationship between statistical significance and publication probability for these data types.[50]

4.5 Public and Private Data

Another meaningful distinction is between data sources that have a public-use version versus those that do not (i.e., private data). Following our conceptual framework which focuses on the threat of a potential p-hacker being detected, we might expect research using public data sources to be less p-hacked than research using private data (given the increase of the probability of a later independent researcher accessing the data). However, we do note two facts: (1) public datasets often differ from private in their purpose and underlying variables (in the same manner third-party and administrative dataset might) and so the research using public and private data may not be *directly* comparable and (2) less (more) researchers have access to private (public) datasets, potentially increasing (decreasing) the marginal contribution of research based on it. We note that if publication bias is a combination of both statistical significance *and* contribution to the literature, it could also differently affect research derived from public and private data sources.

Our sample consists of 1,845 test statistics that rely solely on private data sources, and 15,613 tests that only use data collected by public entities. We omit articles using own-collected data or a mix of own-collected, public and private data.[51]

third-party survey data. We provide derounded and article weighted results for our calipers in Table A18 and A19.

[49] We describe the methodology more fully in Section 3.

[50] For both admin and own-collected data, a two-star result is more likely to be published than a three-star result by not-insignificant margin, while for third-party survey and other data types, a two-star result is almost as likely to be published as a three-star result (97.5% and 94.4%, respectively.

[51] Private data that becomes public through being shared due to (for example) a later data and

We first check whether it is harder to share private data (or harder to obtain for other researchers) than it is for public data. In our full sample around one-third of all data can be directly accessed. Among studies relying solely on private data, about 5% of test statistics provide direct access to the data and 83% provide no access nor help in obtaining the data. For studies using public data, 17% of test statistics provide direct access, while 59% provide no access nor help in obtaining the data.

We now turn to comparing the distribution of test statistics for studies using solely private or public data. Appendix Figure A35 provides this comparison. Visually, we find that the spike near the 5% significance threshold is more visible for private data. We formally examine this possibility using caliper tests. The results for the 5%, 10% and 1% statistical significance levels are presented in Appendix Tables A23-A25, respectively. These tables have the same structure as Table 9, but the key independent variable is a dummy variable that takes the value one if the dataset is categorized as 'public' and zero if it is categorized as 'private'. The point estimates are all statistically insignificant, suggesting no significant differences in p-hacking and publication bias for public and private-data users.

5 Role of the Review Process

In this section, we investigate the role of the reviewing process in mitigating or exacerbating the extent of p-hacking by journal data policy and data type. For this exercise, we directly compare the distribution of z-statistics in our sample of published articles to the distribution of z-statistics in the corresponding working papers for each data type. The objective of this exercise is to document whether journal editors and reviewers require or propose changes that would lead to meaningful changes in the prevalence of marginally significant tests.

For this analysis, we focus on the sample from Brodeur et al. (2020). In order to document the impact of the reviewing process, we only rely on working papers released before the date of submission to the journal. For the 11 journals for which we do not have the date of submission, we rely only working papers released at least two years prior to publication. For those with multiple working papers, we chose the working paper closest to the date of submission (or the two-year threshold). Our final sample of working papers comprises 133 articles/working papers published in 2015 and 2018.[52]

Our data collection methodology for the working papers is the same as for the published version. While some working papers include additional main tests/tables,

code policy, we continue to classify it as private.

[52]The likelihood to find a valid working paper is not statistically related to data type. See Appendix Table A26.

or rely on different clustering or weighting techniques, we find that the distribution of z-statistics is remarkably similar between the working paper and published version for journals with a data-sharing policy. Appendix Figure A36 illustrates the estimated kernel densities for the working paper and published version, respectively. The curves for the working paper and published version are mostly on top of each other for the entire distribution. Similarly, Appendix Figure A37 shows the estimated kernel densities for the working paper and published version for journals without a data-sharing policy. Again, the curves are on top of each other, suggesting limited changes to the main results from the submission to the publication.

For data types, we also find that the distribution of z-statistics is remarkably similar between the working paper and published version for all data types. Appendix Figure A38 illustrates the estimated kernel densities for the working paper and published version for all data types. Again, the curves are on top of each other for the entire distribution for all data types.

We formally test whether there are changes in reporting of significant results due to the reviewing process using the following equation:

$$Pr(Significant_{i,a} = 1) = \Phi(\alpha + \omega + \nu_a) \tag{4}$$

where $Significant_{i,a}$ is an indicator variable that test statistic i is statistically significant for for the 5% significance threshold. Following Brodeur et al. (2020), we define ω as one when test statistic i is in the working-paper version of article a and zero otherwise. Article fixed effects are represented by ν_a. We apply this parsimonious equation to the entire sample and then restrict to subsamples, allowing for flexible estimation of α.

Appendix Table A27 reports the caliper tests. Column 1 includes all data types while columns 2–5 restrict the sample to admin, third-party survey, own-collected and other data type, respectively. Overall, we find that the estimated effect of the publication process is very small and statistically insignificant for the entire sample and for each type separately. This leads us to believe the editorial process does not change the extent of selective reporting.

6 Conclusion

Demands for and use of 'big data' to analyze different aspects of our lives, our society, and our economy continue to grow. In this paper, we documented unexplored facets of the link between data-sharing policies, methodologies for data collection and research transparency; the extent of p-hacking across data types and whether data-sharing policies decrease p-hacking. Our analysis points to no appreciable effect of data-sharing policies or between-data type differences.

These results are key from the point of view of journals and their editors considering the implementation of a data-sharing policy. Of note, our results should not be viewed as indicating that archives for replication material are not useful.[53] Our results instead suggest that the benefits of such archives are more limited than previously thought, but still extremely valuable by allowing, for instance, other researchers to replicate research findings. Our results are also key for policymakers and researchers who are interested in knowing to what extent they should be skeptical about the credibility of the published literature using specific data types. Overall our findings suggest that the increased monitoring (perceived or real) resulting from providing data and code is not a key factor driving p-hacking in economics.

To conclude, we briefly discuss some of the limitations of our study. First, our study deals with journal articles from top economics journals, and thus our findings might not generalize to less elite academic outlets. Second, we are unable to say much about the colloquial 'file drawer' problem (where researchers abandon projects to the file drawer after finding, for example, statistically insignificant results), and cannot say much of the possibility that studies using specific data types are more or less likely to remain unpublished. This could be an issue if unpublished studies that are more/less p-hack are more/less likely to use a specific methodology for data collection.

[53] It is worth mentioning the recent increase in formal restricted-access data environments, which facilitate access to admin data for a large number of researchers. Examples of such environments include the U.S. Federal Statistical Research Data Center and the German IAB FDZ.

References

Abadie, A.: 2020, Statistical Nonsignificance in Empirical Economics, *American Economic Review: Insights* **2**(2), 193–208.

Adda, J., Decker, C. and Ottaviani, M.: 2020, P-hacking in Clinical Trials and How Incentives Shape the Distribution of Results across Phases, *Proceedings of the National Academy of Sciences* **117**(24), 13386–13392.

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

Becker, G. S.: 1968, Crime and punishment: An economic approach, *The Economic Dimensions of Crime*, Springer, pp. 13–68.

Blanco-Perez, C. and Brodeur, A.: 2019, Transparency in Empirical Economic Research, *IZA World of Labor* p. 467.

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

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A. and Olds, J. L.: 2015, Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. Report of the Subcommittee on Replicability in Science Advisory Committee.

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

Bruns, S. B., Asanov, I., Bode, R., Dunger, M. et al.: 2019, Reporting Errors and Biases in Published Empirical Findings: Evidence from Innovation Research, *Research Policy* **48**(9), 103796.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T. et al.: 2016, Evaluating Replicability of Laboratory Experiments in Economics, *Science* **351**(6280), 1433–1436.

Card, D. and DellaVigna, S.: 2013, Nine Facts about Top Journals in Economics, *Journal of Economic Literature* **51**(1), 144–61.

Card, D. and DellaVigna, S.: 2020, What do Editors Maximize? Evidence from Four Economics Journals, *Review of Economics and Statistics* **102**(1), 195–217.

Card, D., DellaVigna, S., Funk, P. and Iriberri, N.: 2020, Are Referees and Editors in Economics Gender Neutral?, *Quarterly Journal of Economics* **135**(1), 269–327.

Carrell, S., Figlio, D. and Lusher, L.: 2020, Clubs and Networks in Economics Reviewing. working paper.

Cattaneo, M. D., Jansson, M. and Ma, X.: 2020, Simple Local Polynomial Density Estimators, *Journal of the American Statistical Association* **115**(531), 1449–1455.

Cattaneo, M. D., Jansson, M. and Ma, X.: 2021, rddensity: Manipulation Testing Based on Density Discontinuity in R. Working Paper, University of Michigan.

Chang, A. C., Li, P. et al.: 2022, Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not", *Critical Finance Review* **11**(1), 185–206.

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.

DellaVigna, S. and Linos, E.: 2022, RCTs to Scale: Comprehensive Evidence from Two Nudge Units, *Econometrica* **90**(1), 81–116.

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.

Einav, L. and Levin, J.: 2014, Economics in the Age of Big Data, *Science* **346**(6210), 1243089.

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

Feige, E. L.: 1975, The Consequences of Journal Editorial Policies and a Suggestion for Revision, *Journal of Political Economy* **83**(6), 1291–1296.

Franco, A., Malhotra, N. and Simonovits, G.: 2014, Publication Bias in the Social Sciences: Unlocking the File Drawer, *Science* **345**(6203), 1502–1505.

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

Gerber, A. and Malhotra, N.: 2008a, 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.

Gerber, A., Malhotra, N. et al.: 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.

Gerber, A. S. and Malhotra, N.: 2008b, Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?, *Sociological Methods & Research* **37**(1), 3–30.

Hamermesh, D. S.: 2017, Replication in Labor Economics: Evidence from Data, and What it Suggests, *American Economic Review* **107**(5), 37–40.

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.

Havránek, T., Stanley, T., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W. R., Rost, K. and Van Aert, R.: 2020, Reporting Guidelines for Meta-Analysis in Economics, *Journal of Economic Surveys* **34**(3), 469–475.

Höffler, J. H.: 2017, Replication and Economics Journal Policies, *American Economic Review: Papers and Proceedings* **107**(5), 52–55.

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

Kapteyn, A. and Ypma, J. Y.: 2007, Measurement Error and Misclassification: A Comparison of Survey and Administrative Data, *Journal of Labor Economics* **25**(3), 513–551.

Künn, S.: 2015, The Challenges of Linking Survey and Administrative Data, *IZA World of Labor* .

Maniadis, Z., Tufano, F. and List, J. A.: 2017, To Replicate or not to Replicate? Exploring Reproducibility in Economics Through the Lens of a Model and a Pilot Study, *Economic Journal* **127**(605).

McCullough, B., McGeary, K. A. and Harrison, T. D.: 2008, Do Economics Journal Archives Promote Replicable Research?, *Canadian Journal of Economics* **41**(4), 1406–1420.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D., Humphreys, M., Imbens, G. et al.: 2014, Promoting Transparency in Social Science Research, *Science* **343**(6166), 30–31.

Mueller-Langer, F., Fecher, B., Harhoff, D. and Wagner, G. G.: 2019, Replication Studies in Economics—How Many and Which Papers Are Chosen for Replication, and Why?, *Research Policy* **48**(1), 62–83.

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

Simonsohn, U., Nelson, L. D. and Simmons, J. P.: 2014, P-curve: A Key to the File-Drawer, *Journal of Experimental Psychology: General* **143**(2), 534.

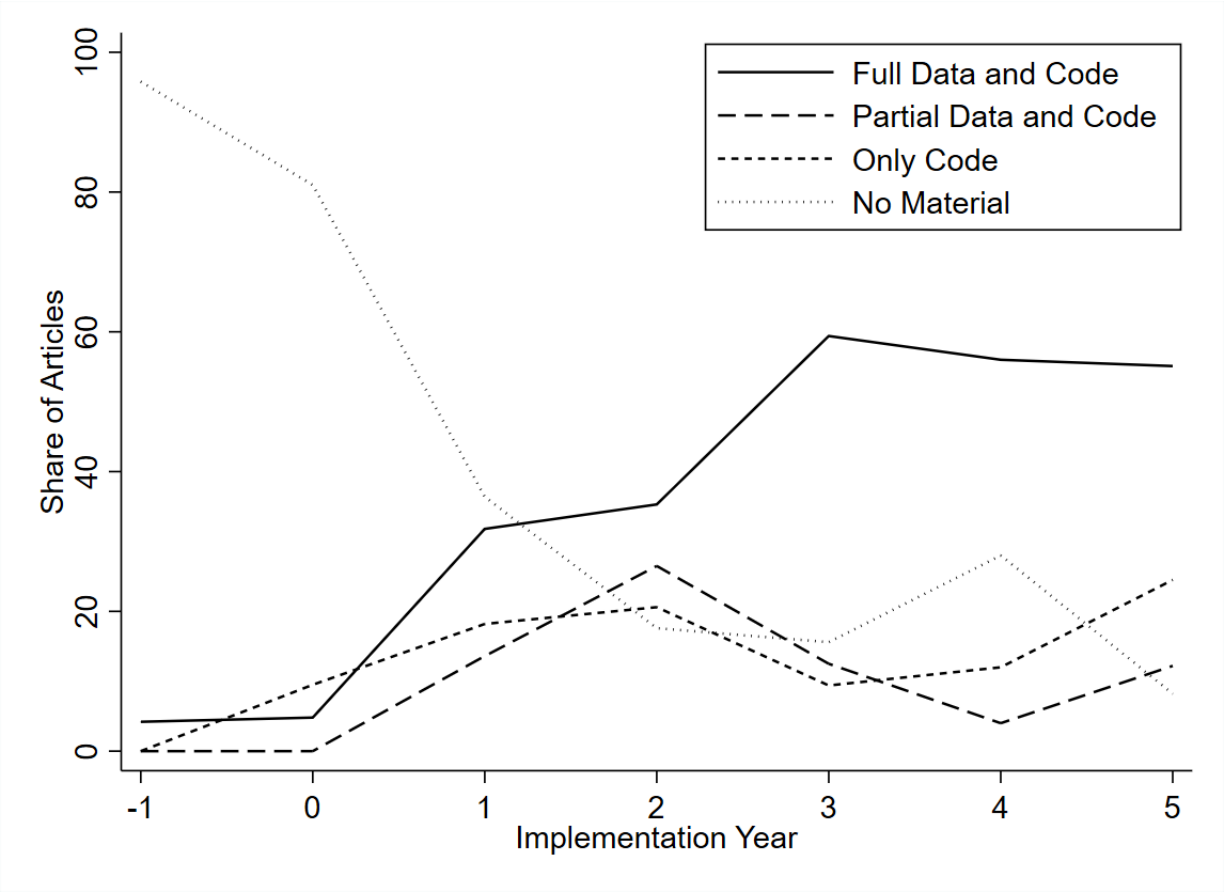
Stanley, T. D.: 2008, Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection, *Oxford Bulletin of Economics and Statistics* **70**(1), 103–127.

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

Swanson, N., Christensen, G., Littman, R., Birke, D., Miguel, E., Paluck, E. L. and Wang, Z.: 2020, Research Transparency Is on the Rise in Economics, *AEA Papers and Proceedings*, Vol. 110, pp. 61–65.

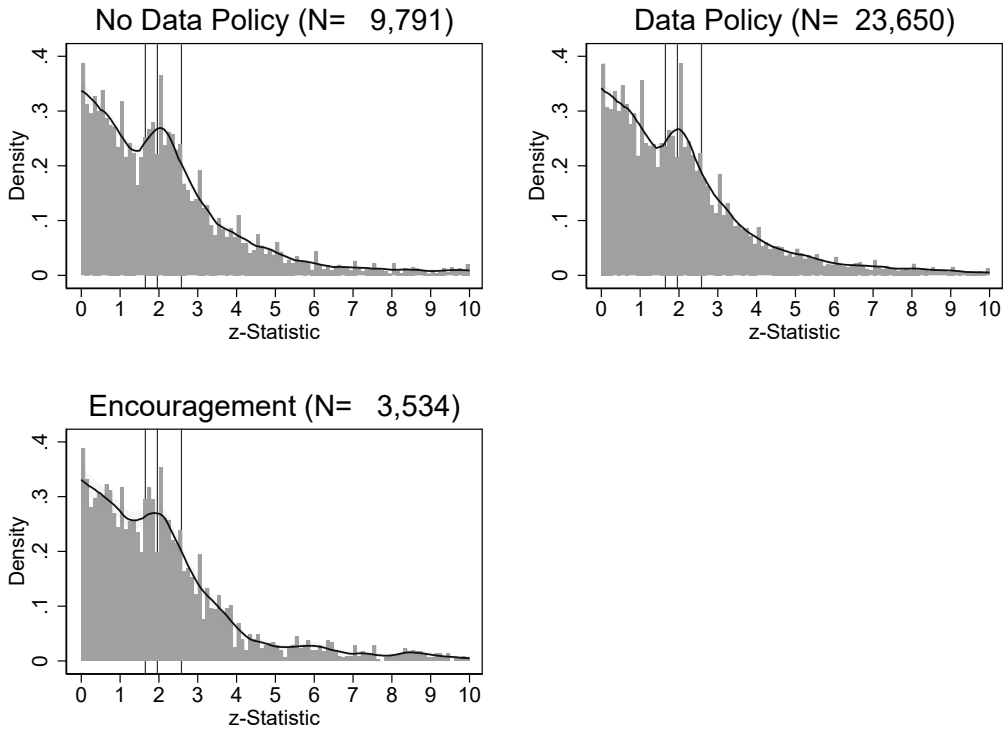
Vivalt, E.: 2019, Specification Searching and Significance Inflation Across Time, Methods and Disciplines, *Oxford Bulletin of Economics and Statistics* **81**(4), 797–816.

Figure 1: Data and Code Availability by Implementation Year



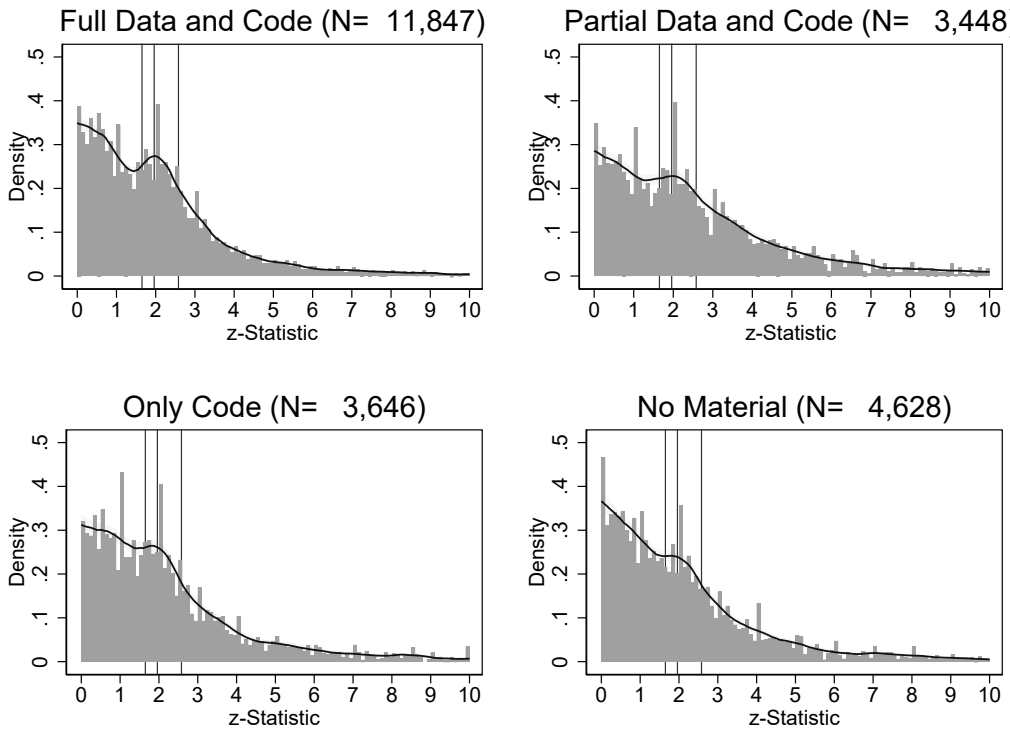
Notes: This figure displays the provision of data and code by implementation year for: *Full Provision of Data and Code*, *Partial Data and Code*, *Only Code*, and *No Material*. We consider a balanced sample of journals: *American Economic Review*, *Journal of Political Economy*, *Econometrica*, *Quarterly Journal of Economics*, *Review of Economic Studies*, *Review of Economics and Statistics*, *Journal of the European Economic Association*, *Economic Journal*, and *Journal of Labor Economics*. Of note, one article published in the *Quarterly Journal of Economics* released full data and codes on the journal's website prior to the implementation of the data-sharing policy.

Figure 2: Estimated Density of z-Statistics by Journal Policy



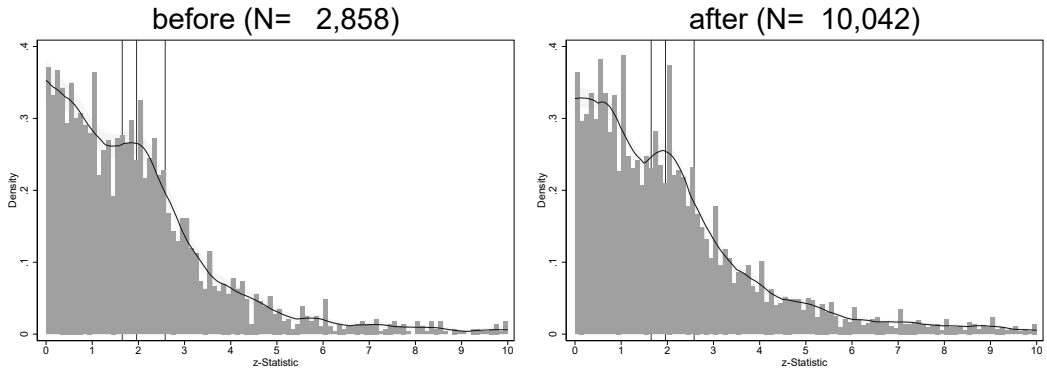
Notes: This figure displays 36,975 test statistics for $z \in [0, 10]$. See Table 1 for an overview of journals by journal policy. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure 3: Estimated Density of z-Statistics for Journals with Data-Sharing Policy by Data and Code Availability



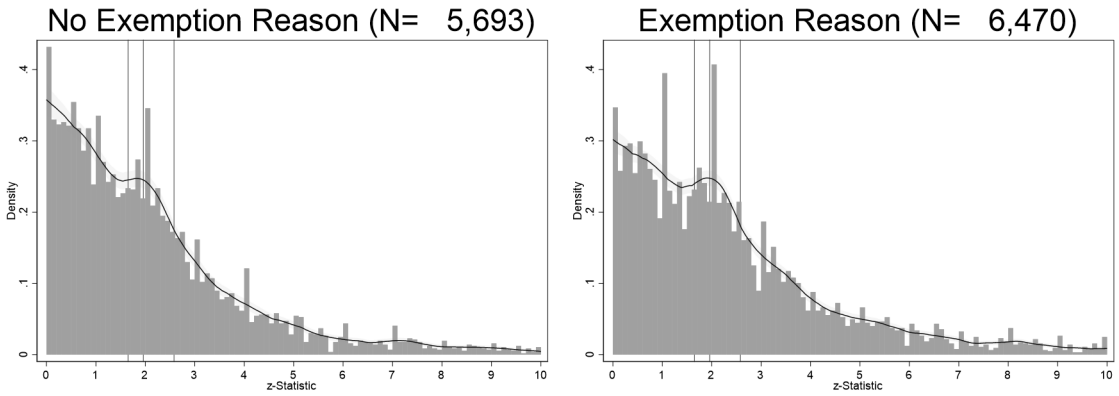
Notes: This figure restricts the sample to those tests that have a Journal policy and displays 23,650 test statistics for $z \in [0, 10]$. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure 4: Estimated Density of z-Statistics and Data-Sharing Policy: Five Years Prior vs. Five Years After



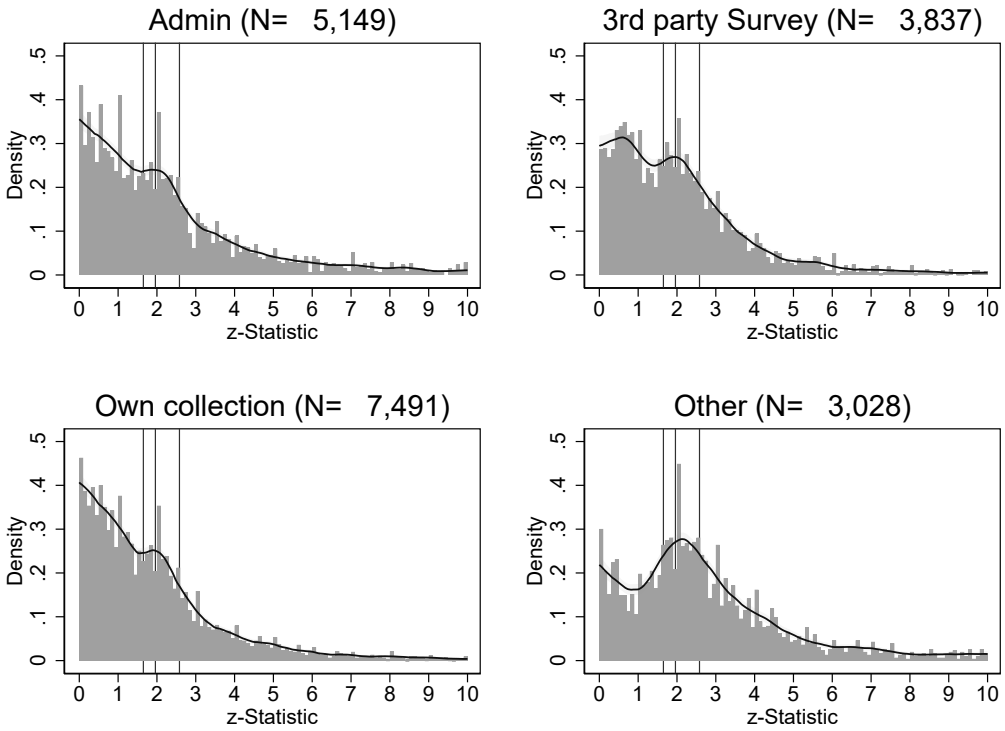
Notes: This figure restricts the sample to observations five years before and five years after a policy change. It displays 12,900 test statistics for $z \in [0, 10]$. See Table 1 for an overview of journals by journal policy. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure 5: Estimated Density of z-Statistics for Partial and No Data Provision: Reason for Data Exemption



Notes: We restrict the sample to articles published in journals with a data-sharing policy and have partial or no data. In our sample, we have 24,742 tests with a journal policy and 12,747 of those tests do not provide full data and code. 53.3% or 6,791 test statistics provide a reason for data exemption, while 46.7% do not. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure 6: Estimated Density of z-Statistics by Data Type



Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by data type: *admin*, *third-party survey*, *own-collected* and *other*. We restrict the sample to studies using only one method of data collection. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Table 1: Journal Data-Sharing Policies

Journal	Policy	Announcement (Year)	# Articles	# Test Statistics	Data Collection (Year)
American Economic Review	Yes	2004	132	5,238	2002-2020
A. Econ. J.: Applied Econ.	Yes	2009	50	2,470	2015, 2018
A. Econ. J.: Econ. Policy	Yes	2009	42	1,251	2015, 2018
A. Econ. J.: Macroeconomics	Yes	2009	5	54	2015, 2018
Econometrica	Yes	2004	22	578	2002-2020
Economic Journal	Yes	2012	78	2,629	2002-2020
Economic Policy	Yes	2017	6	2,629	2015, 2018
Experimental Economics	Encourage		6	79	2015, 2018
J. of Applied Econometrics	Yes	1994	5	86	2015, 2018
J. of Development Economics	Yes	2014	64	2,818	2015, 2018
J. of Economic Growth	Encourage		8	100	2015, 2018
Journal of Finance	Only Code	2018	51	2,084	2002-2020
J. of Financial Economics	No		39	569	2015, 2018
J. of Finan. Intermediation	Encourage		16	185	2015, 2018
J. of Human Resources	Yes	2019	57	1,697	2002-2020
J. of International Econ.	No		19	488	2015, 2018
J. of Labor Economics	Yes	2010	39	1,114	2002-2020
J. of Political Economy	Yes	2005	51	1,854	2002-2020
J. of Public Economics	Encourage		74	2,605	2015, 2018
J. of Urban Economics	Encourage		26	660	2015, 2018
J. of the Euro. Econ. Ass.	Yes	2011	56	1,648	2002-2020
Quarterly Journal of Econ.	Yes	2016	71	3,951	2002-2020
Review of Economic Studies	Yes	2006	26	1,634	2002-2020
Review of Econ. & Stat.	Yes	2010	96	3,286	2002-2020
Review of Financial Studies	No		67	1,618	2002-2020
Total			1106	38876	

Notes: This table provides an overview of data and code journal policies. We obtained this information on the journals' websites. We also obtained some of the implementation dates from Christensen and Miguel (2018) and Mueller-Langer et al. (2019). Columns 3 and 4 report the number of articles and test statistics in our sample. Column 5 reports the years for which we collected test statistics.

Table 2: Summary Statistics: Data and Code Availability and Author Characteristics

	Availability of Replication Material				Total
	Full Data and Code	Partial Data and Code	Only Code	No Material	
	(1)	(2)	(3)	(4)	(5)
Top 5	0.51	0.37	0.37	0.22	0.34
	(0.50)	(0.48)	(0.48)	(0.41)	(0.47)
Editor present	0.56	0.51	0.33	0.47	0.49
	(0.50)	(0.50)	(0.47)	(0.50)	(0.50)
Solo-authored	0.10	0.14	0.09	0.12	0.11
	(0.30)	(0.34)	(0.29)	(0.32)	(0.31)
Average	12.22	9.71	10.44	10.40	10.93
experience	(6.47)	(5.01)	(6.44)	(5.90)	(6.14)
Female authors	0.20	0.23	0.17	0.21	0.20
	(0.28)	(0.28)	(0.28)	(0.29)	(0.29)
Top institutions	0.27	0.26	0.23	0.20	0.23
	(0.30)	(0.34)	(0.30)	(0.26)	(0.29)
Top PhD	0.37	0.37	0.30	0.29	0.33
institutions	(0.31)	(0.35)	(0.32)	(0.31)	(0.32)
Test statistics	12580	3828	3954	18514	38876

Notes: Each observation is a test. 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*. Average experience is the mean of years since PhD for an article's authors. The other variables are the share of female authors, share of authors affiliated with top institutions, and share of authors who completed a PhD at a top institution, respectively.

Table 3: Summary Statistics: Data Type and Article and Author Characteristics

	Method of Data Collection				Total
	Admin	3rd Party Survey	Own Collected	Other	
	(1)	(2)	(3)	(4)	(5)
Top 5	0.39	0.21	0.40	0.25	0.34
	(0.49)	(0.41)	(0.49)	(0.44)	(0.47)
Editor present	0.43	0.38	0.67	0.50	0.52
	(0.49)	(0.48)	(0.47)	(0.50)	(0.50)
Sole authored	0.17	0.17	0.08	0.09	0.12
	(0.38)	(0.38)	(0.27)	(0.28)	(0.33)
Average	10.34	10.16	13.15	12.43	11.69
experience	(7.20)	(6.07)	(6.38)	(6.96)	(6.79)
Female authors	0.16	0.24	0.28	0.11	0.21
	(0.29)	(0.34)	(0.28)	(0.22)	(0.29)
Top institutions	0.28	0.11	0.29	0.22	0.24
	(0.31)	(0.20)	(0.32)	(0.29)	(0.30)
Top PhD	0.30	0.23	0.44	0.26	0.33
institutions	(0.33)	(0.25)	(0.35)	(0.28)	(0.33)
Test statistics	5726	3932	7754	3289	20701

Notes: Each observation is a test. In this table, we only consider those observations that rely solely on one data type within each study ("pure" sample). 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*. Average experience is the mean of years since PhD for an article's authors. The other variables are the share of female authors, share of authors affiliated with top institutions, and share of authors who completed a PhD at a top institution, respectively.

Table 4: Caliper Tests 5% Threshold: Data-Sharing

	Significant at 5% Level					
	(1)	(2)	(3)	(4)	(5)	(6)
Data & Code Provided (Author)	-0.038*	-0.020	-0.042*	-0.031	-0.026	-0.021
	(0.020)	(0.021)	(0.023)	(0.023)	(0.028)	(0.029)
Data & Code Provided (Journal)	0.008	0.019	0.006	0.016	0.010	0.020
	(0.019)	(0.019)	(0.020)	(0.020)	(0.024)	(0.024)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	9,334	9,332	6,829	6,828	4,060	4,059
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variables take a value one if the test statistic is drawn from an article with data and code provided (on the author's homepage or journal's website) and zero otherwise.

Table 5: Caliper Tests 5% Threshold: Data-Sharing and Data-Sharing Policy

	(1)	(2)	(3)	(4)	(5)	(6)
First Stage			DV: Data and Code Provided			
Data & Code Policy	0.797***	0.788***	0.731***	0.675***	0.725***	0.697***
	(0.167)	(0.173)	(0.173)	(0.160)	(0.173)	(0.171)
Second Stage			DV: Significant at 5% Level			
Data & Code Provided	-0.253	-0.503	-0.070	-0.239	0.291	0.266
(Journal)	(0.387)	(0.436)	(0.456)	(0.501)	(0.514)	(0.555)
Reduced Form			DV: Significant at 5% Level			
Data & Code Policy	-0.046	-0.080	-0.011	-0.031	0.048	0.037
	(0.066)	(0.064)	(0.072)	(0.066)	(0.084)	(0.076)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	3,177	3,177	2,282	2,282	1,362	1,362
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: **All Panels:** Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. **First Stage:** Requiring data and code to be posted increases the provision of data and code. The dependent variable takes a value one if the test statistic is drawn from an article with data and code provided on the journal's website and zero otherwise. The primary independent variable takes a value one if the article was published in a journal up to 5 years after the journal implemented a data and code availability requirement and zero up to 5 years before the journal implements a policy. **Second Stage:** Provision of data and code, instrumented by journal policy, does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variable takes a value one if the article was predicted to provide data and code on the journal's website using journal policy as an instrument. **Reduced Form:** Requiring data and code to be posted does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variable takes a value one if the article is published in a journal with a data and code policy.

Table 6: Results from Application of Elliott et al. (2022): Data and Code Availability

Sample	Bin.	Disc.	CS1	CS2B	LCM	Obs	Total
Full Sample							
Full Data And Code	0.000	0.002	0.002	0.000	0.006	419	7368
Other	0.000	0.760	0.000	0.000	0.000	840	15929
Pre-Journal Policy							
Full Data And Code	.	.	.	.	.	1	103
Other	0.008	0.863	0.140	0.000	0.794	83	1657
Post-Journal Policy							
Full Data and Code	0.006	0.082	0.002	0.000	0.179	145	2428
Other	0.000	0.001	0.000	0.000	0.418	186	3682
(Only Journals That Switch)							
Full Data And Code	0.005	0.029	0.002	0.000	0.185	146	2531
Other	0.000	0.005	0.008	0.000	0.140	269	5339

Notes: This table provides the result from the battery of tests proposed in Elliott et al. (2022) for the 5% threshold for data and code availability.

Table 7: Relative Publication Probabilities by Data and Code Availability

Sample	u	t	df	[0,1.1645]	(1.645,1.960]	(1.960,2.576]
All	0.003	0.003	1.486	0.357	0.791	1.134
Full Sample						
Full Data and Code	0.003	0.003	1.659	0.340	0.769	1.121
Other	0.002	0.003	1.419	0.360	0.787	1.130
Pre-Journal Policy						
Full Data and Code	0.062	0.078	2.177	0.558	1.173	1.509
Other	0.005	0.007	1.366	0.636	1.335	1.584
Post-Journal Policy						
Full Data and Code	0.003	0.002	1.666	0.350	0.807	1.123
Other	0.001	0.003	1.470	0.346	0.658	0.987
(Only Journals That Switch)						
Full Data and Code	0.003	0.002	1.655	0.337	0.786	1.101
Other	0.003	0.004	1.434	0.418	0.826	1.140

Notes: The table presents the results of applying the publication bias model presented in Andrews and Kasy (2019) to data and code availability. The model assumes that the underlying effect sizes follow a generalized t distribution. We report the model's estimated location parameter, scale parameter, and degrees of freedom in the first three columns. In the fourth column, 0.357 represents the relative probability that a test statistic in the [0,1.1645] interval is 35.7% as likely to be published as a test statistic greater than 2.576 (the reference interval).

Table 8: Prediction of Provision of Full Data and at Least Code

Provision of ...	Full Data and Code			At least Code		
	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
Third-Party Survey	0.230***	0.297***	0.255***	0.025	0.128*	0.170
	(0.064)	(0.062)	(0.059)	(0.089)	(0.076)	(0.283)
Own Collection	0.530***	0.420***	0.421***	0.220***	0.202**	0.727***
	(0.061)	(0.062)	(0.057)	(0.080)	(0.084)	(0.273)
Other	0.182**	0.153***	0.236***	-0.032	-0.017	0.557**
	(0.084)	(0.052)	(0.061)	(0.094)	(0.067)	(0.271)
Controls						
DID		-0.038	-0.031		0.080	0.297
		(0.055)	(0.055)		(0.072)	(0.247)
IV		-0.065	-0.073		0.066	0.224
		(0.050)	(0.051)		(0.072)	(0.257)
RDD		– 0.189**	– 0.191**		0.042	0.160
		(0.076)	(0.076)		(0.092)	(0.348)
Top 5		0.285***	1.262***		0.398***	7.090***
		(0.040)	(0.087)		(0.053)	(0.185)
Experience		0.012	0.012		0.007	0.024
		(0.009)	(0.008)		(0.011)	(0.042)
Experience^2		-0.024	-0.023		-0.011	-0.027
		(0.024)	(0.022)		(0.030)	(0.111)
Top Institution		-0.045	-0.025		0.019	0.048
		(0.081)	(0.073)		(0.101)	(0.325)
PhD Top Institution		-0.098	-0.101		0.001	0.009
		(0.075)	(0.070)		(0.090)	(0.286)
Other Controls						
Year FE	Y	Y	Y	Y	Y	Y
Solo Authored		Y	Y		Y	Y
Share Female Authors		Y	Y		Y	Y
Editor		Y	Y		Y	Y
Field			Y			Y
Observations	20,701	19,593	19,222	20,701	20,373	20,002

Notes: We rely on probit models and present the average marginal effects (equation (3)). The dependent variable in column (1)-(3) is a dummy for whether full data and code can be accessed, while the dependent variable is a dummy for whether at least code is available on webpages of the journals for column (4)-(6). The omitted category is *admin*. The omitted category for the methods is RCT. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

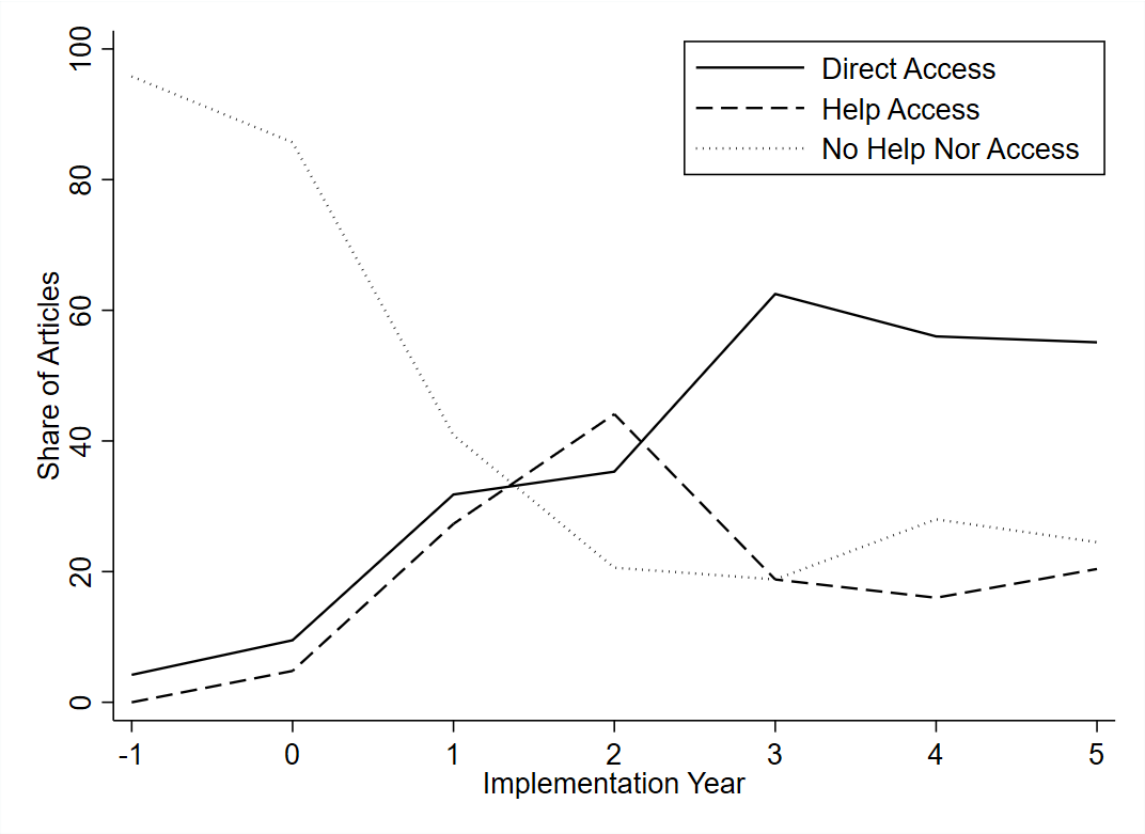
Table 9: Caliper Tests 5% Threshold: Data Type

	Significant at 5% Level					
	(1)	(2)	(3)	(4)	(5)	(6)
Own-Collected	-0.007	0.043	0.010	0.071*	0.021	0.023
	(0.024)	(0.038)	(0.029)	(0.043)	(0.037)	(0.054)
Third-Party Survey	0.002	0.003	-0.009	0.007	0.008	-0.001
	(0.030)	(0.028)	(0.033)	(0.031)	(0.043)	(0.042)
Other	-0.037	-0.051	-0.025	-0.031	-0.009	-0.014
	(0.033)	(0.032)	(0.037)	(0.037)	(0.048)	(0.047)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Data and Code Provided		✓		✓		✓
Additional Controls		✓		✓		✓
Observations	4,814	4,814	3,517	3,517	2,089	2,089
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variables are dummy variables for each data type.

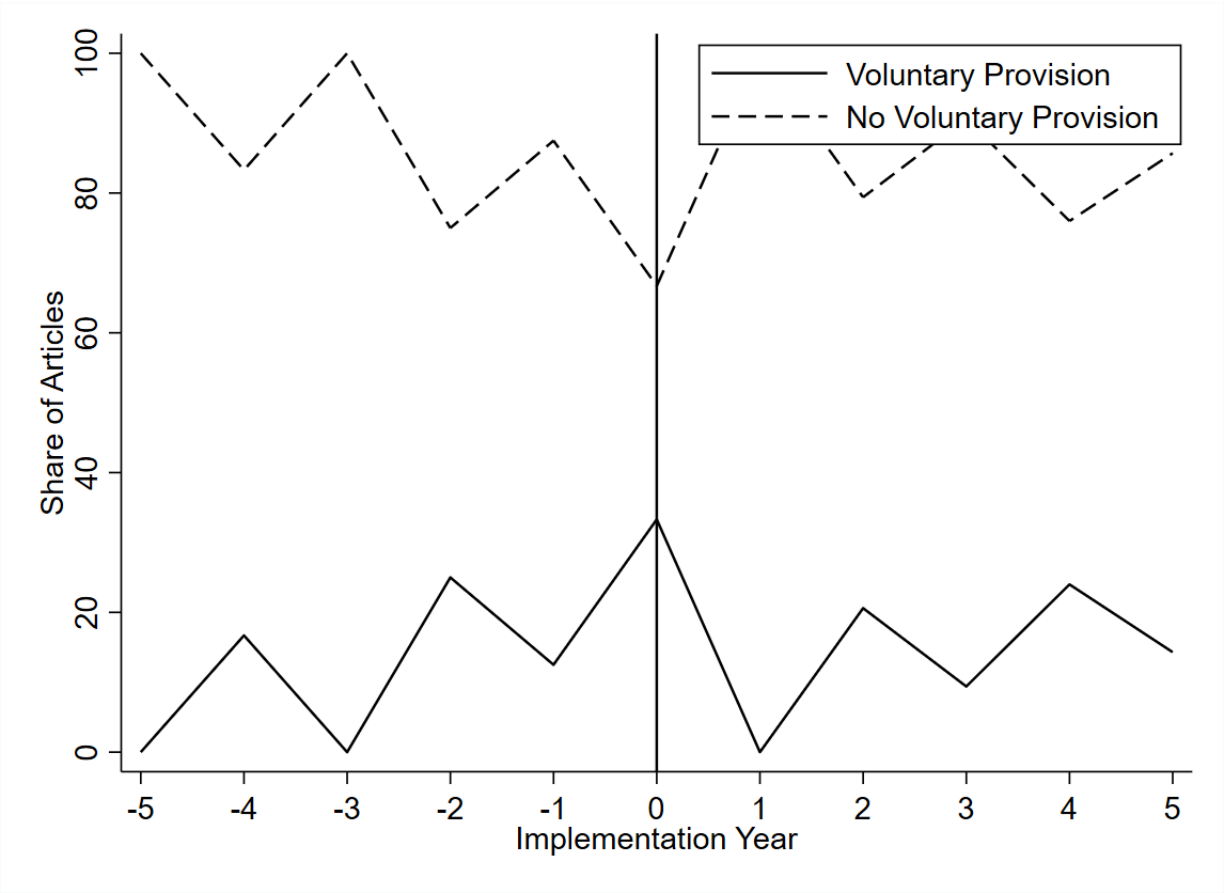
Appendix Figures: NOT FOR PUBLICATION

Figure A1: Data Access by Implementation Year



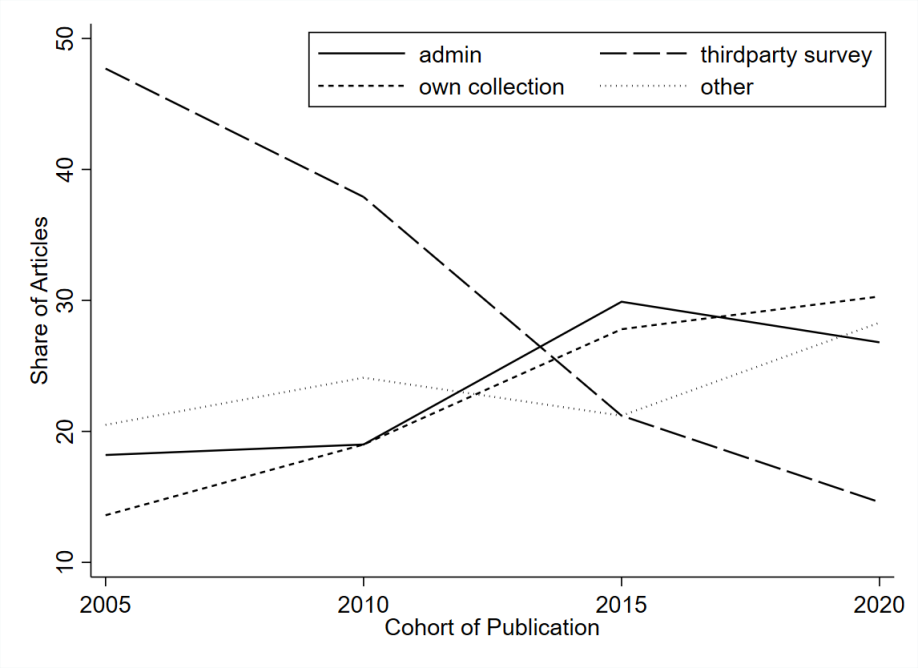
Notes: This figure displays data access by implementation year for: *Direct Access*, *Help from the Authors*, and *No Access*. We consider a balanced sample of journals: *American Economic Review*, *Journal of Political Economy*, *Econometrica*, *Quarterly Journal of Economics*, *Review of Economic Studies*, *Review of Economics and Statistics*, *Journal of the European Economic Association*, *Economic Journal*, and *Journal of Labor Economics*. Of note, one article published in the *Quarterly Journal of Economics* released full data and codes on the journal's website prior to the implementation of the data-sharing policy.

Figure A2: Data and Code Availability on Authors' Homepages by Implementation Year



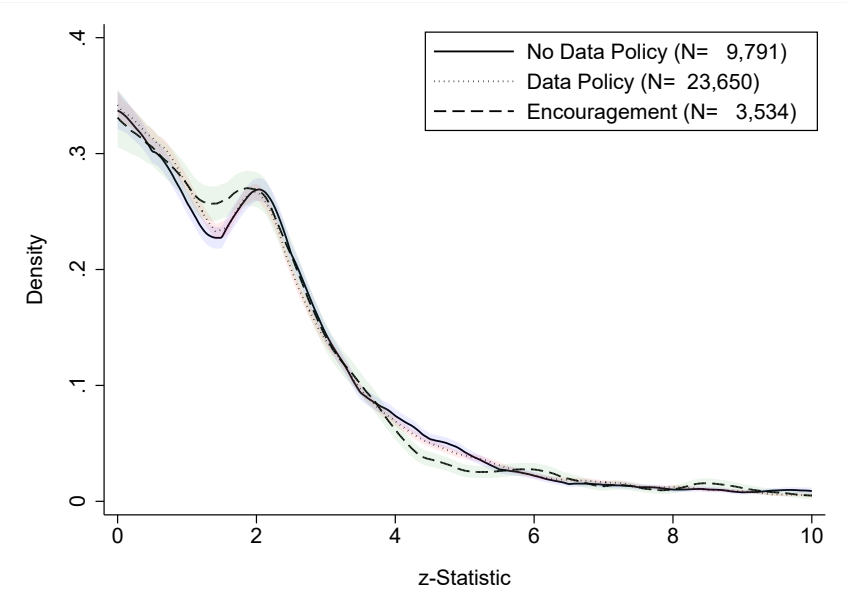
Notes: This figure restricts the sample to observations within five years before and after a policy change. It displays the share of articles that provide data and codes on the authors' homepages. The sample consists of 10,042 test statistics for $z \in [0, 10]$. The solid line represents the share of articles where authors voluntarily upload data and code files on their homepages, while the dashed line show the share of articles without data and code files on authors' homepages.

Figure A3: Data Type by Cohort



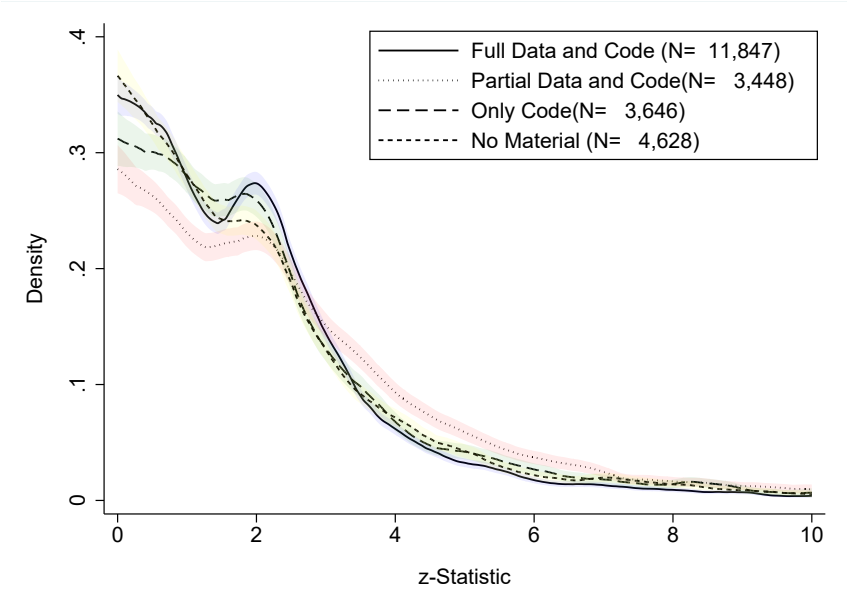
Notes: In this figure, we restrict the sample to articles relying solely on one data type. Cohort 2005 for articles published from 2002–2005; Cohort 2010 for articles published from 2006–2010; Cohort 2015 for articles published from 2011–2015; Cohort 2020 for articles published from 2020 onward.

Figure A4: Estimated Density of z-Statistics by Journal Policy



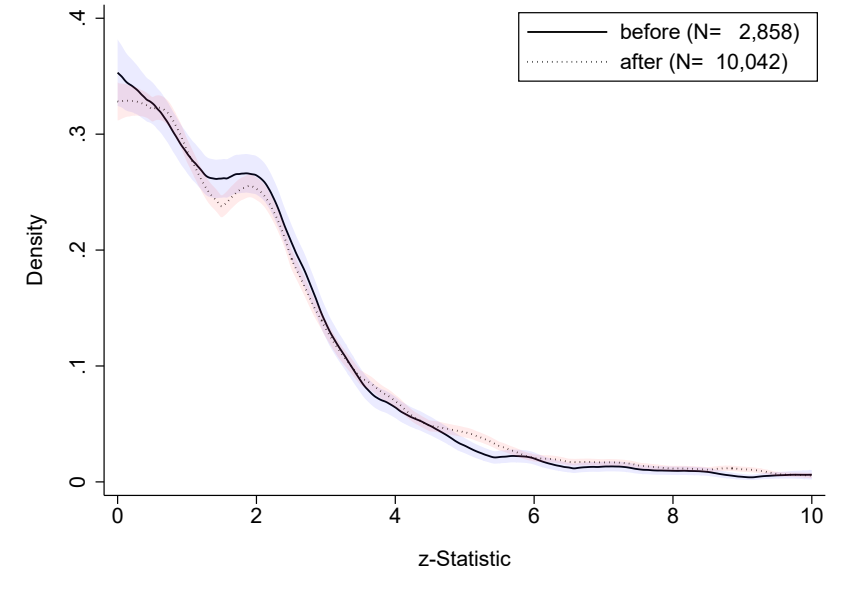
Notes: This figure displays 36,975 test statistics for $z \in [0, 10]$. The solid line presents test statistics with *no data policy*, while the dotted line presents test statistics that face a *journal policy*. The dashed line represents test statistics that do not face strict journal policies but authors are *encouraged* by journals to upload replication files. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates. See Table 1 for an overview of data-sharing policies.

Figure A5: Estimated Density of z-Statistics for Journals with Data-Sharing Policy by Data and Code Availability



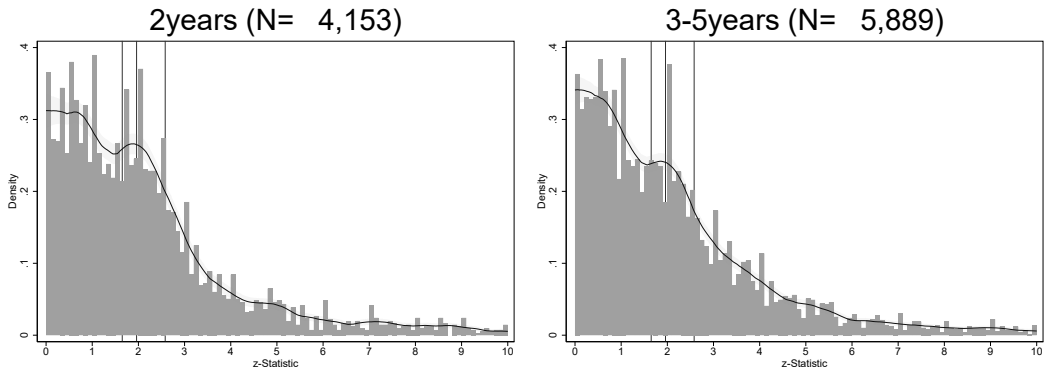
Notes: This figure restricts the sample to those tests that have a Journal policy and displays 23,650 test statistics for $z \in [0, 10]$. The solid line presents test statistics that provide *full data and code*, the dotted line *partial provision of data and code*, the long-dashed line presents tests that provide *only code* and last, the short-dashed line presents those test statistics that provide *no material*. See Table 1 for an overview of data-sharing policies. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A6: Estimated Density of z-Statistics and Data-Sharing Policy: Five Years Prior vs. Five Years After



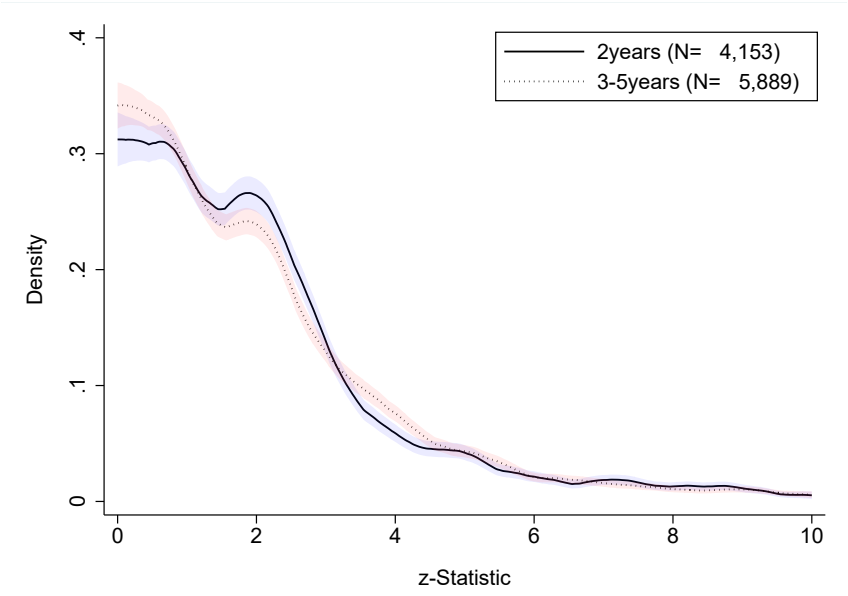
Notes: This figure restricts the sample to observations five years before and five years after a policy change. It displays 12,900 test statistics for $z \in [0, 10]$. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A7: Estimated Density of z-statistics After Data-Sharing Policy: First two Years vs. Three to Five Years



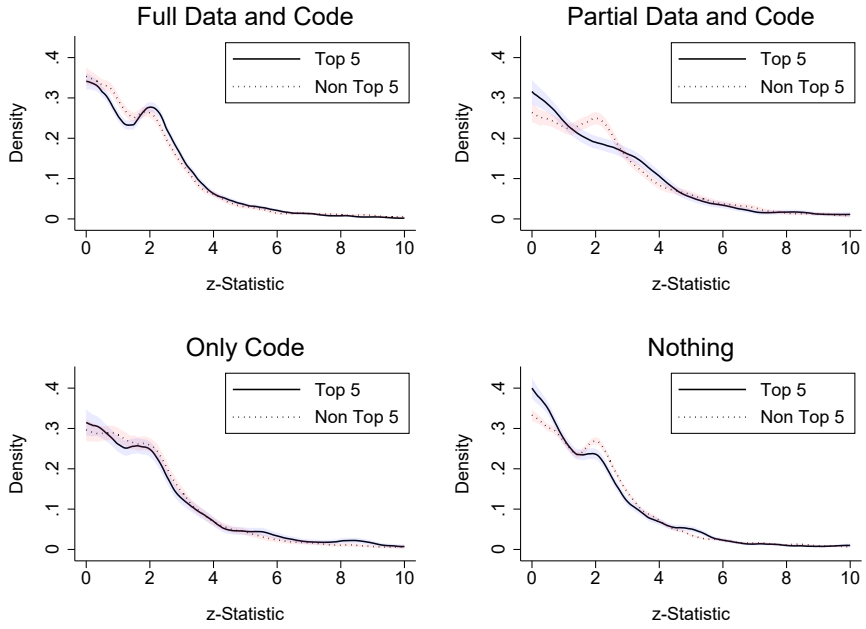
Notes: This figure restricts the sample to observations within five years after a policy change. It displays 10,042 test statistics for $z \in [0, 10]$. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A8: Estimated Density of z-statistics After Data-Sharing Policy: First two Years vs. Three to Five Years



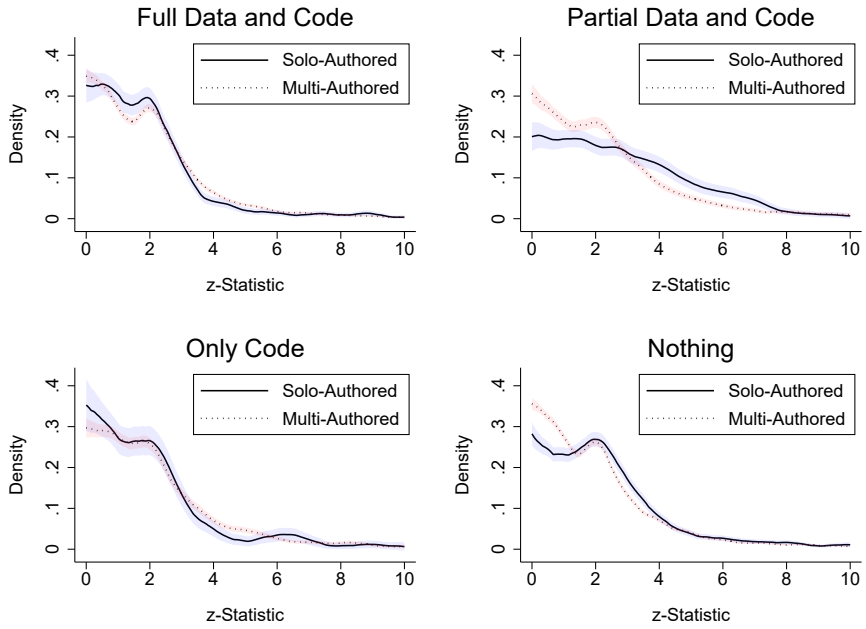
Notes: This figure restricts the sample to observations within five years after a policy change. It displays 10,042 test statistics for $z \in [0, 10]$. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A9: z-statistics by Data and Code Availability and Journal Ranking with Data-Sharing Policy



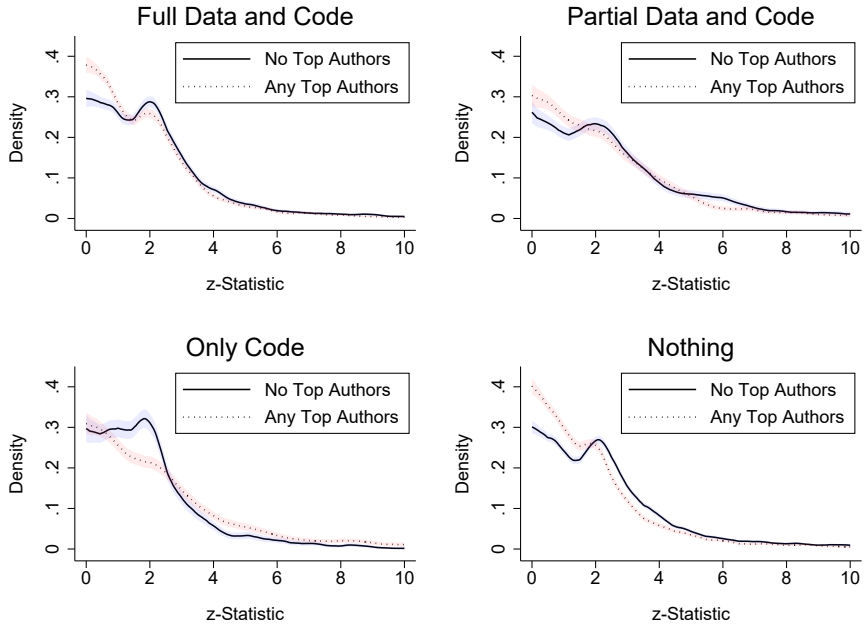
Notes: This figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by data and code availability: Full data and code, partial data and code, only code and nothing. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A10: z-statistics by Data and Code Availability and Number of Authors with Data-Sharing Policy



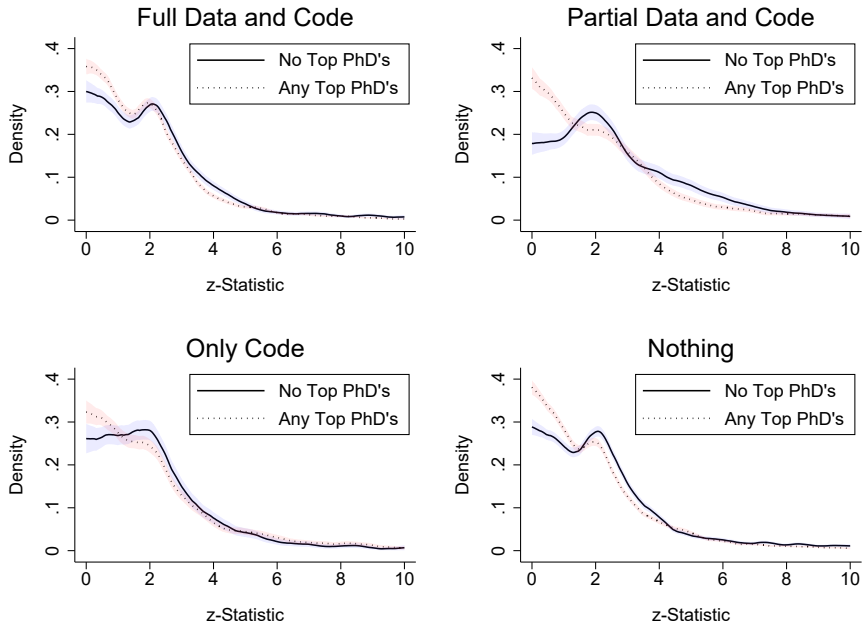
Notes: This figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by data and code availability: Full data and code, partial data and code, only code and nothing. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A11: z-statistics by Data and Code Availability and Affiliation with Data-Sharing Policy



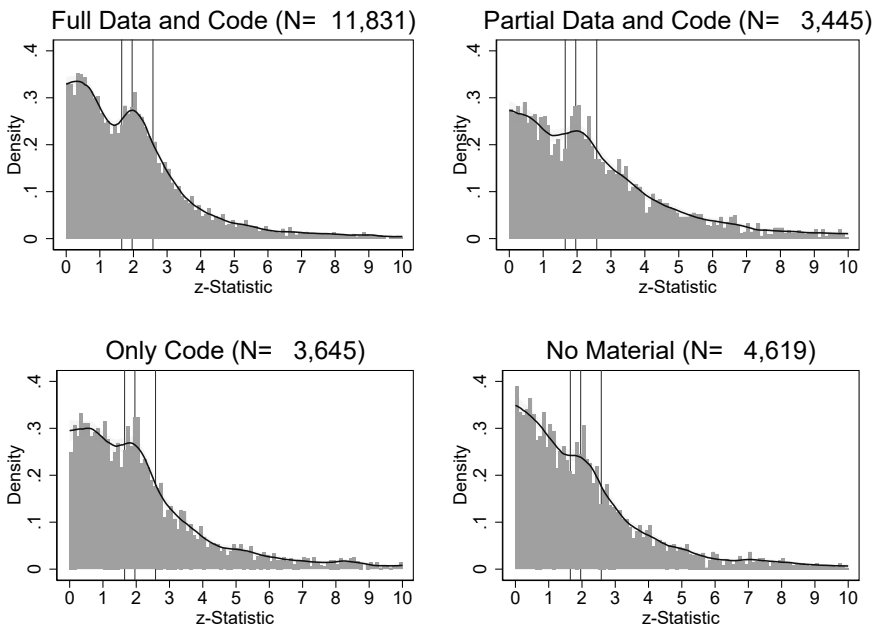
Notes: This figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by data and code availability: Full data and code, partial data and code, only code and nothing. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A12: z-statistics by Data and Code Availability and PhD Institution with Data-Sharing Policy



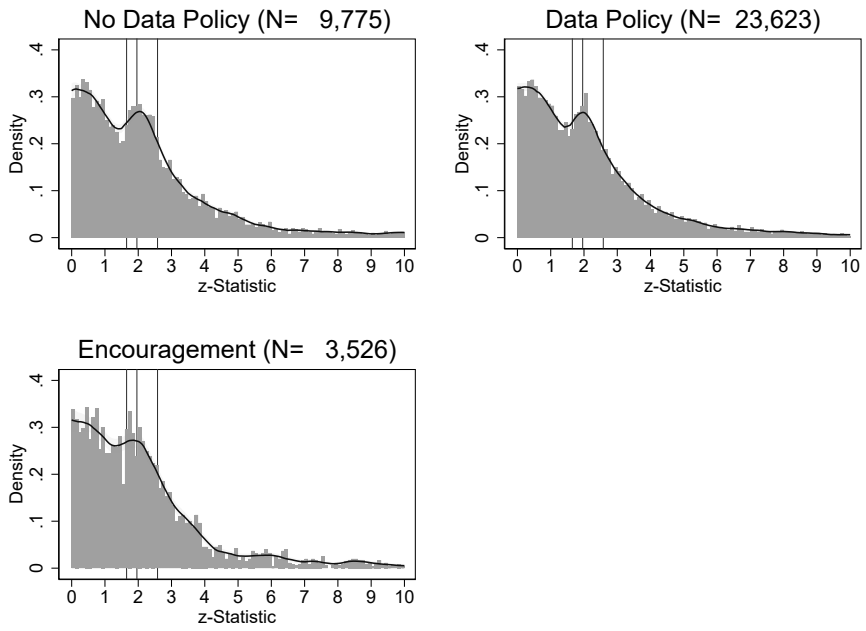
Notes: This figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by data and code availability: Full data and code, partial data and code, only code and nothing. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A13: Estimated Density of z-Statistics for Journals with Data-Sharing Policy by Data Availability (De-Rounded) - with Histograms



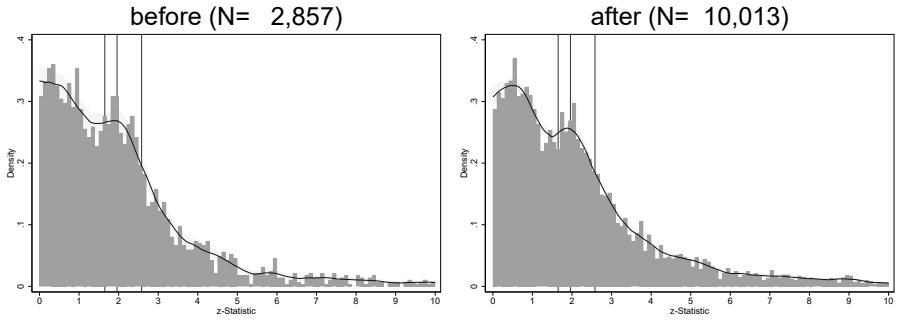
Notes: Notes: This figure restricts the sample to those tests that have a Journal policy and displays 23,542 of test statistics for $z \in [0, 10]$. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. Compared to figure 3 we use de-rounded z-Statistics. We do not weight our estimates.

Figure A14: Estimated Density of z-Statistics by Journal Policy (De-rounded z-Statistics)



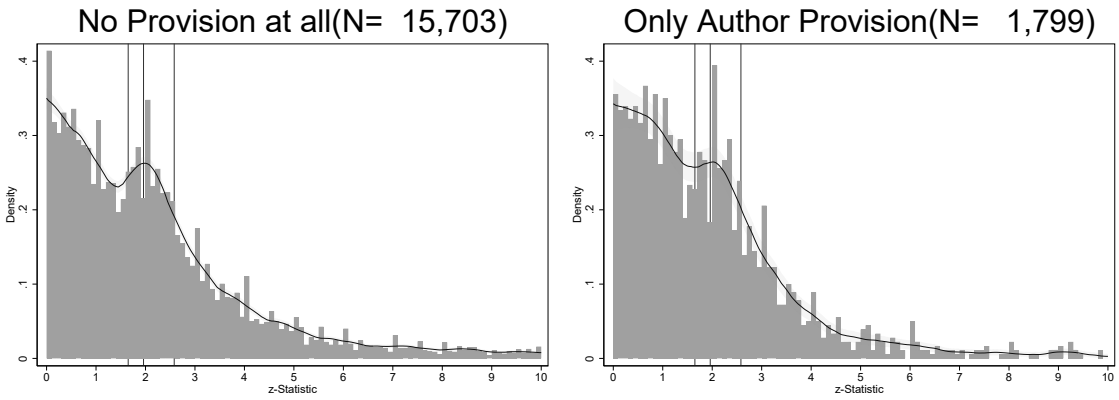
Notes: This figure displays 36,924 of test statistics for $z \in [0, 10]$. See Table 1 for an overview of journals by journal policy. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. Compared to figure 2 we use de-rounded z-Statistics. We do not weight our estimates.

Figure A15: Estimated Density of z-Statistics and Data-Sharing Policy: Five Years Prior vs. Five Years After (De-rounded z-Statistics)



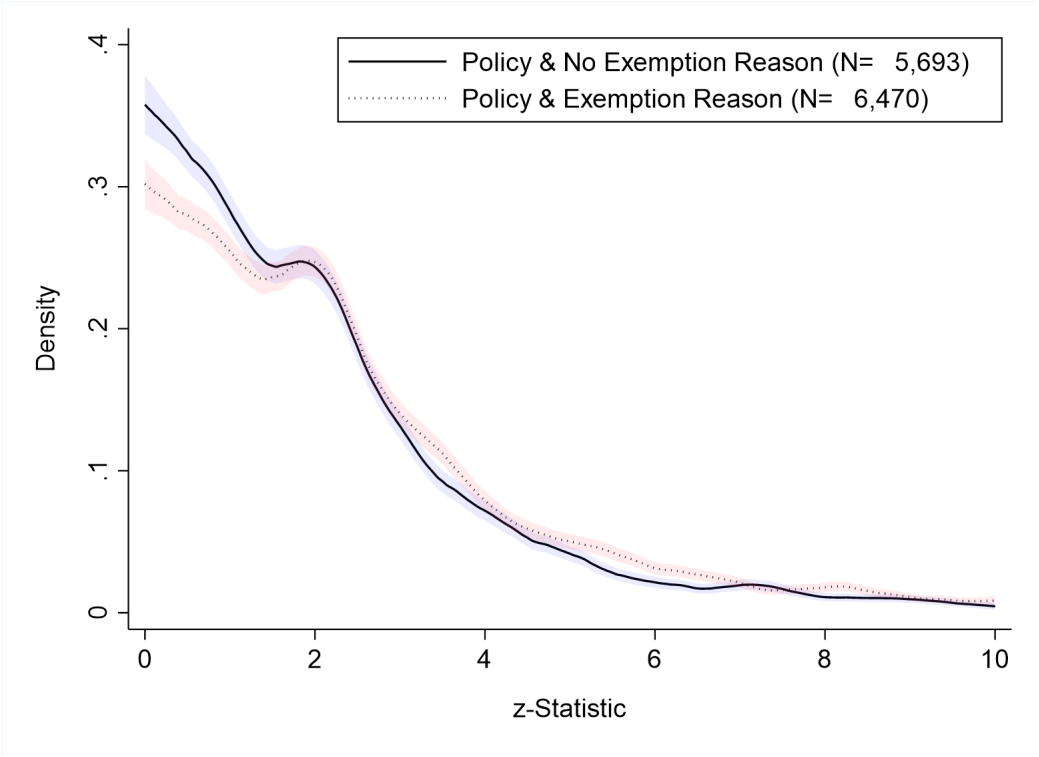
ADJUST Notes: This figure restricts the sample to observations five years before and five years after a policy change. It displays 12,870 of test statistics for $z \in [0, 10]$. See Table 1 for an overview of journals by journal policy. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. Compared to figure 4 we use de-rounded z-Statistics. We do not weight our estimates.

Figure A16: Estimated Density of z-statistics by Data and Code Availability on Personal Homepages



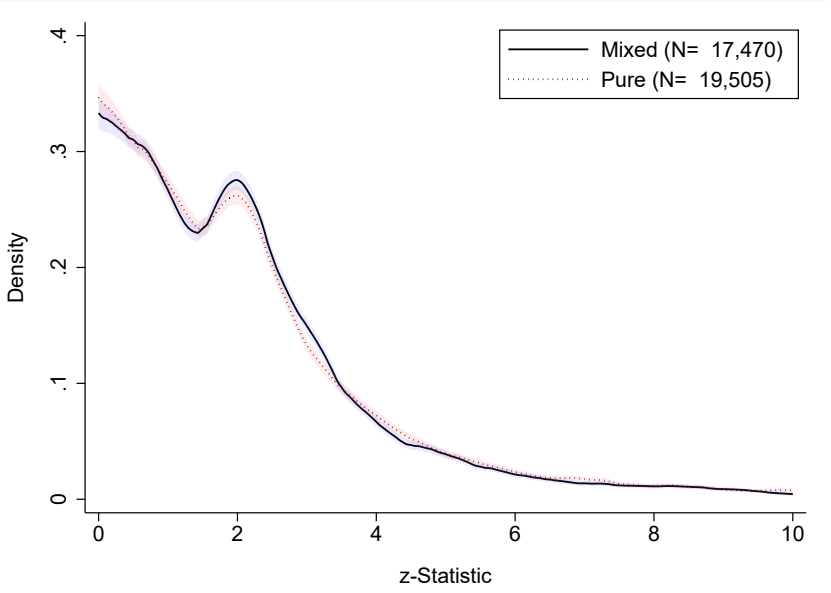
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. The left figure plots all tests with no provision of replication files either on the journal or on authors' homepages. The right figure displays only test statistics with provision only on authors' but not on journals' homepages.

Figure A17: Incomplete Data Provision and Implemented Journal Policy: Reason for Data Exemption



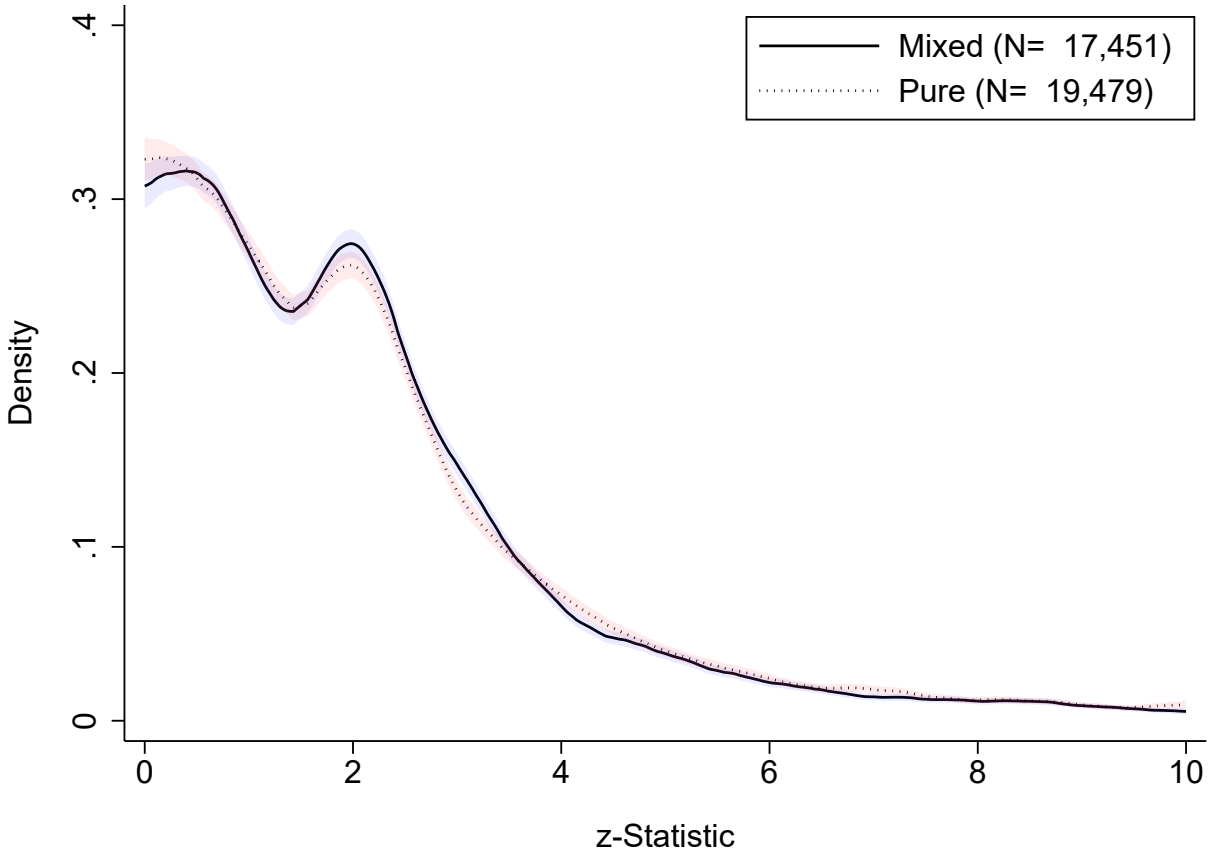
Notes: We restrict the sample to articles published in journals with a data-sharing policy and have partial or no data. In our sample, we have 24,742 tests with a journal policy and 12,747 of those test statistics do not provide full replication material. 53.28% or 6,791 test statistics provide a reason for data exemption, while 46.72% do not. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A18: Estimated Density of z-statistics for Full and Pure Samples



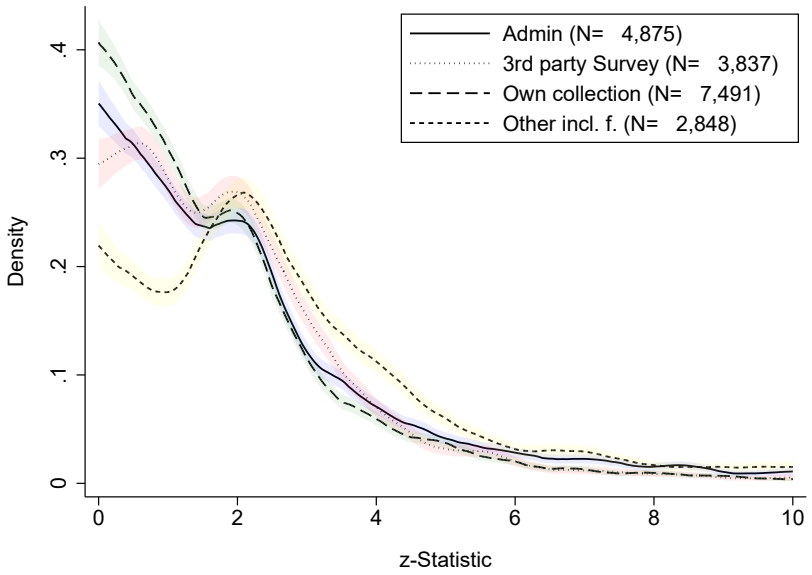
Notes: This figure displays two distributions. First, the solid line plots the z-statistics for those estimates that rely on a 'mixed' data type and second, the dotted line plots the z-statistics for the sub-sample of estimates that rely solely on one data type. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A19: Estimated Density of z-statistics for Full and Pure Samples (De-rounded z-statistics



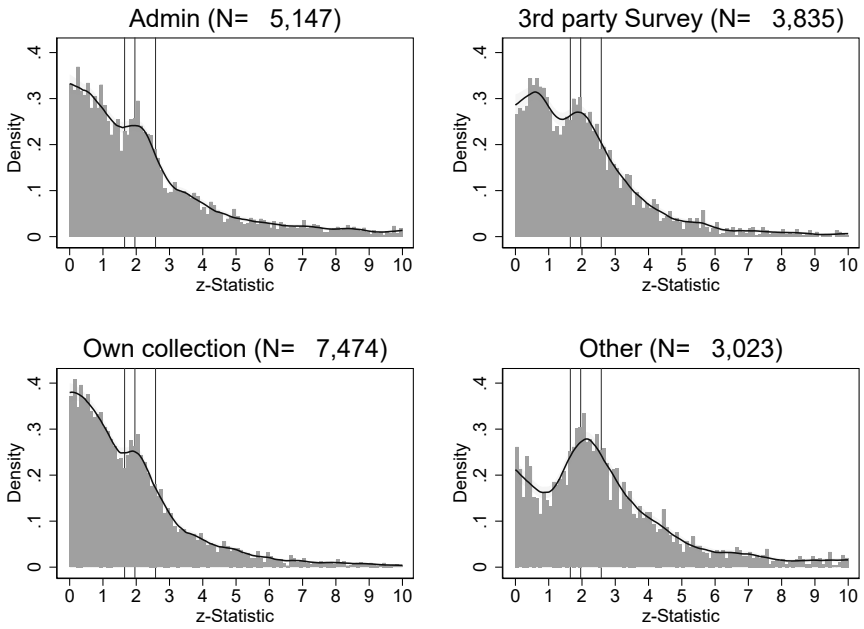
Notes: This figure displays two distributions. First, the solid line plots the z-statistics for those estimates that rely on a 'mixed' data type and second, the dotted line plots the z-statistics for the sub-sample of estimates that rely solely on one data type. Compared to figure A18 we use de-rounded z-statistics.

Figure A20: Estimated Density of z-Statistics by Data Type



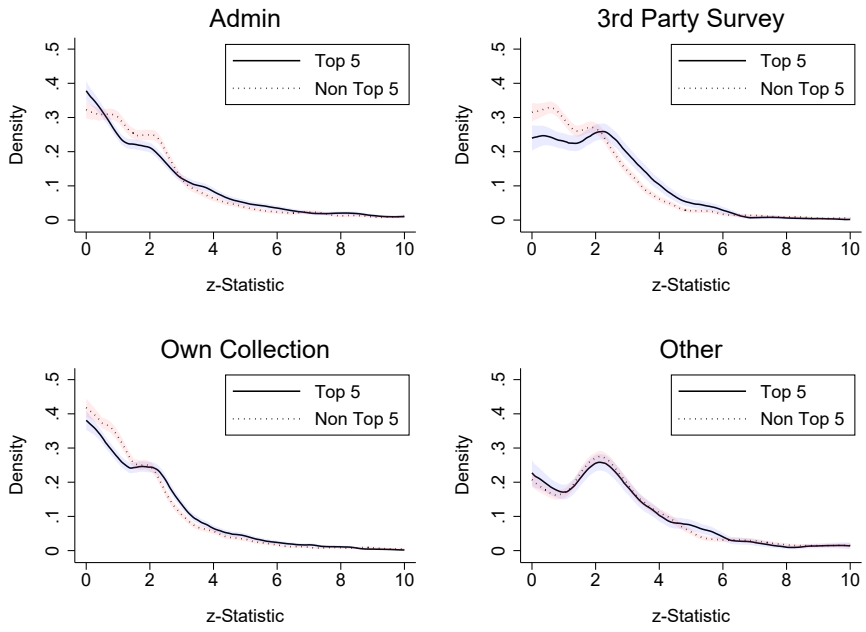
Notes: This figure displays four distributions of test statistics for $z \in [0, 10]$ by data type: *admin*, *third-party survey*, *own-collected* and *other*. We restrict the sample to studies using solely one data type ("pure sample"). We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A21: Estimated Density of z-Statistics by Data Type (De-rounded) - with Histograms



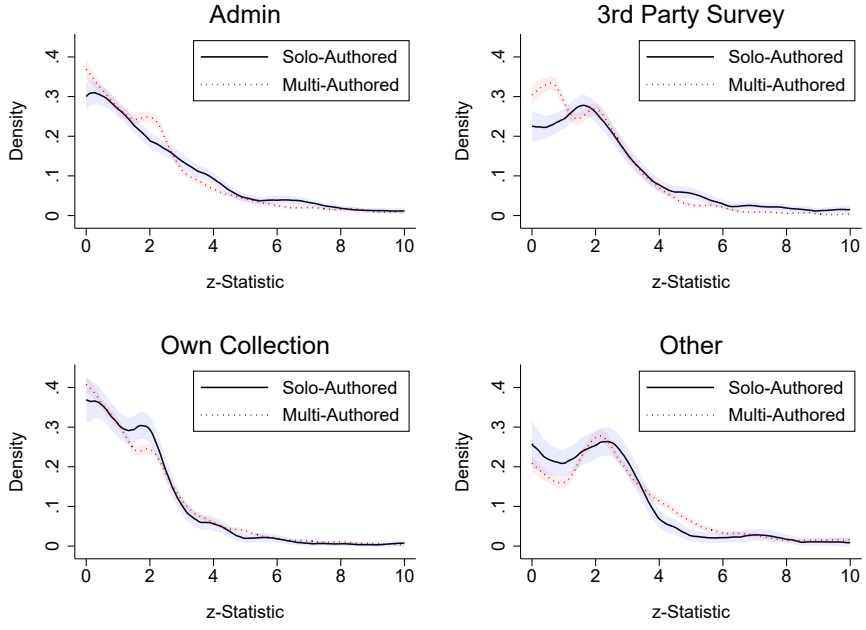
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by data type: *admin*, *third-party survey*, *own-collected* and *other*. We restrict the sample to studies using only one method of data collection. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates. We rely on a de-rounding method (Brodeur et al. (2016)).

Figure A22: Estimated Density of z-statistics by Data Type and Journal Ranking



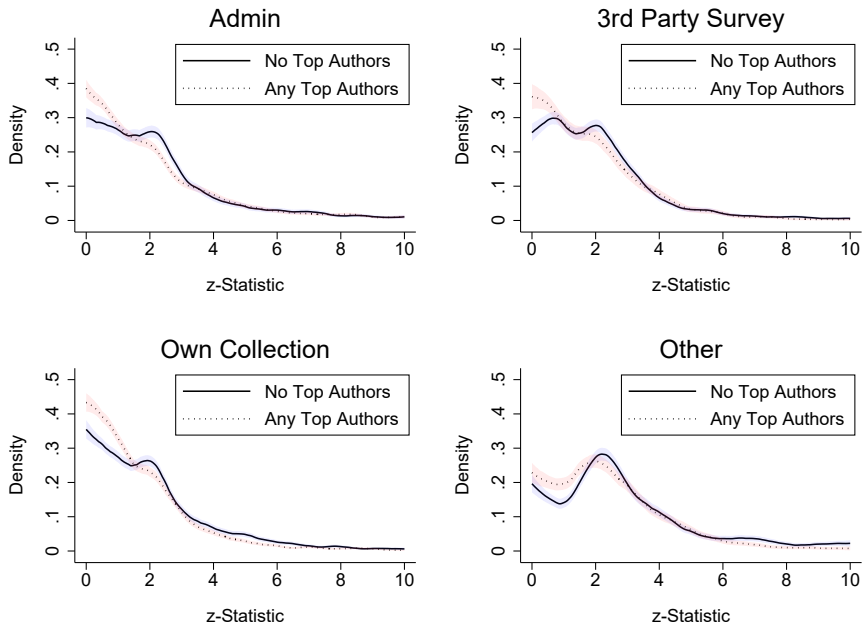
Notes: This figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by data type: Admin, third-party survey, own collection and other. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A23: Estimated Density of z-statistics by Data Type and Number of Authors



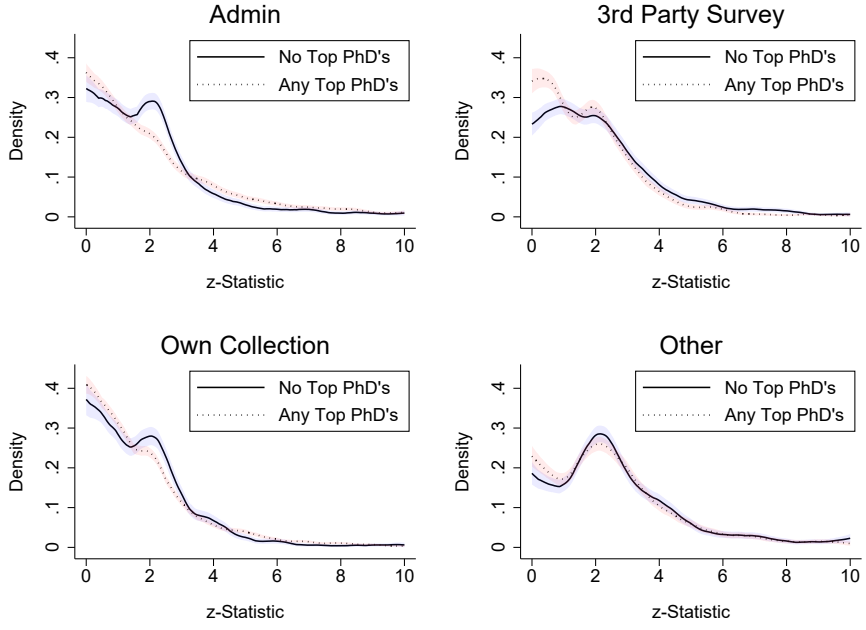
Notes: This figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by data type: admin, third-party survey, own collection and other. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A24: Estimated Density of z-statistics by Data Type and Affiliation



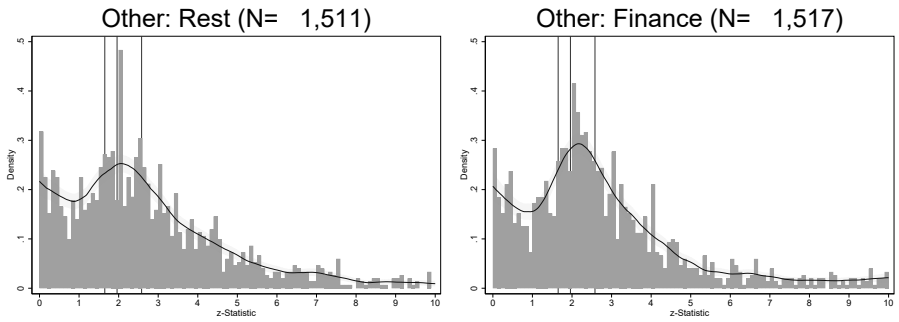
Notes: This figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by data type: admin, third-party survey, own collection and other. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A25: Estimated Density of z-statistics by Data Type and PhD Institution



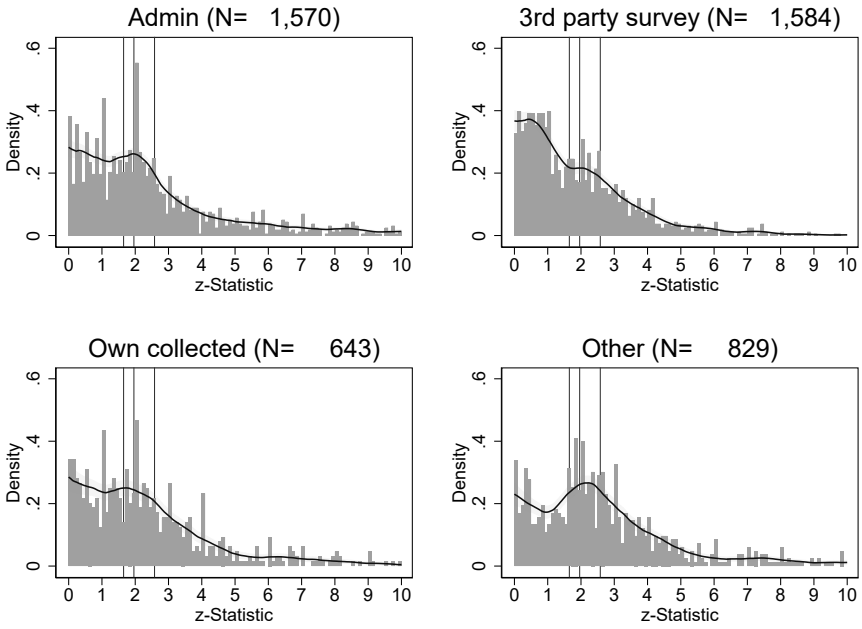
Notes: This figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by data type: admin, third-party survey, own collection and other. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A26: Estimated Density of z-Statistics for Other Category: Financial Data vs Remaining Other



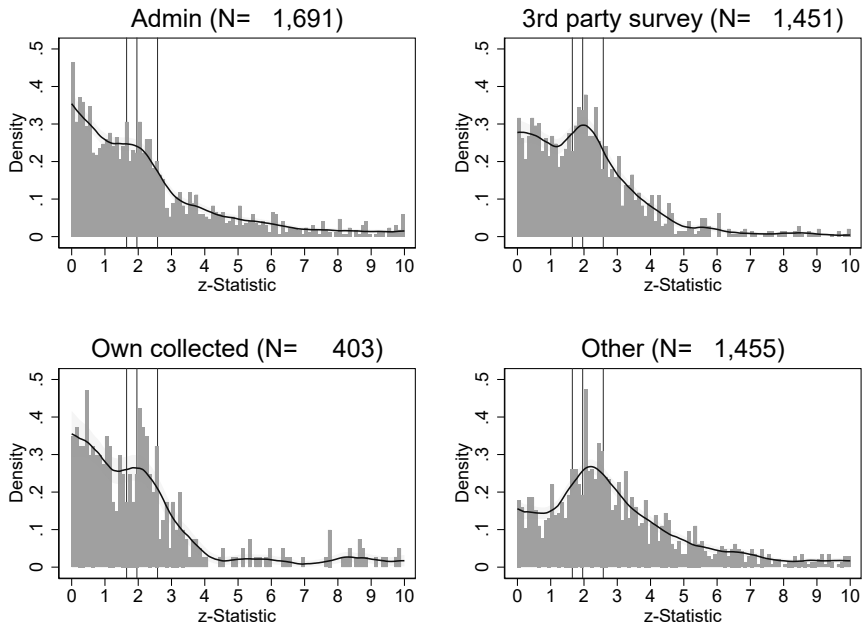
Notes: This figure displays 3,028 of test statistics for $z \in [0, 10]$ only for those that rely on *other data*. As in the pure sample, we consider here only test statistics that uniquely belong to the data type *other*. We split the data type category *other* into those test statistics that use non-financial data (left figure) and those that only rely on financial data (right figure). We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A27: Estimated Density of z-Statistics using DID by Data Type



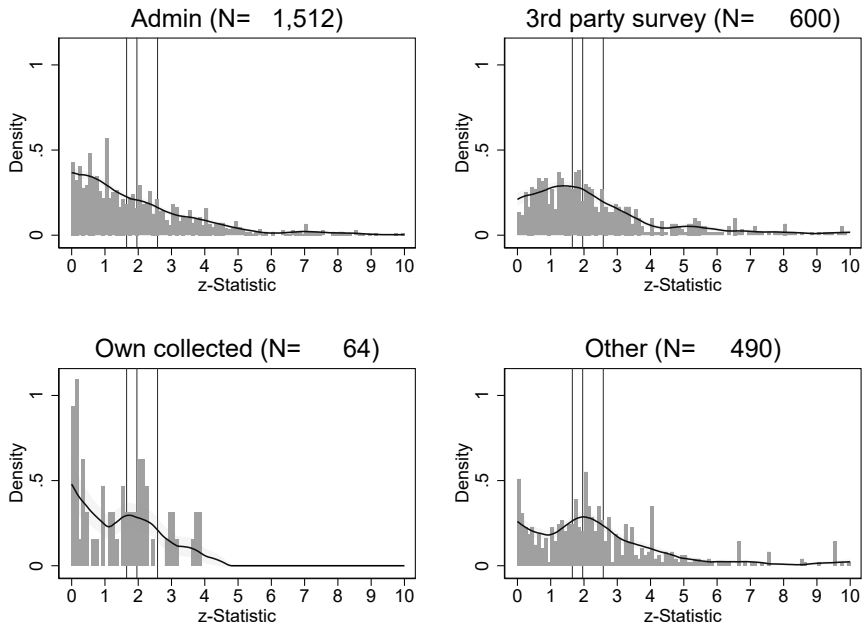
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ using a difference-in-differences (DID) approach by data type: admin, 3rd party survey, own-collected and other. We only consider those observations that rely solely on one type of data within each article. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A28: Estimated Density of z-Statistics using IV by Data Type



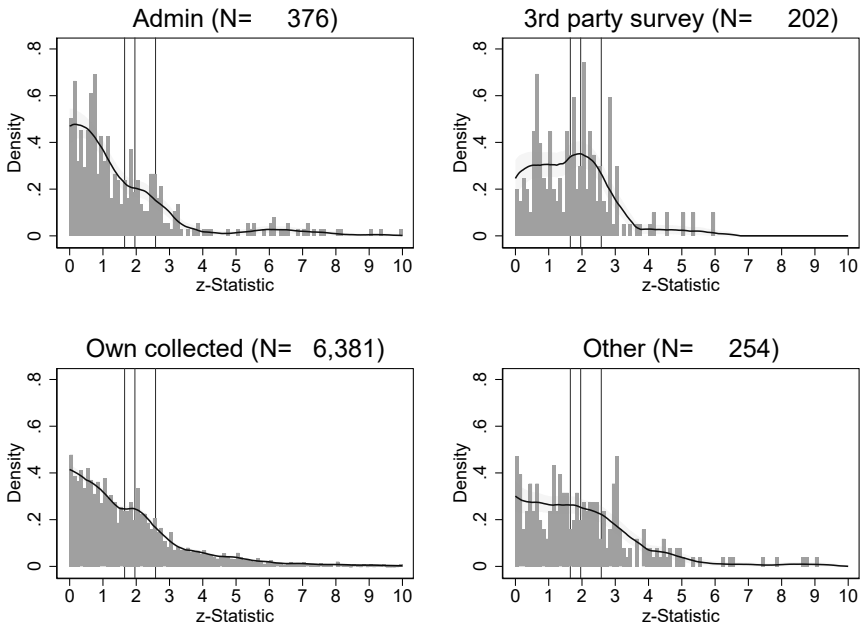
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ using a instrumental variables (IV) approach by data type: admin, third-party survey, own-collected and other. We only consider those observations that rely solely on one type of data within each article. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A29: Estimated Density of z-Statistics using RDD by Data Type



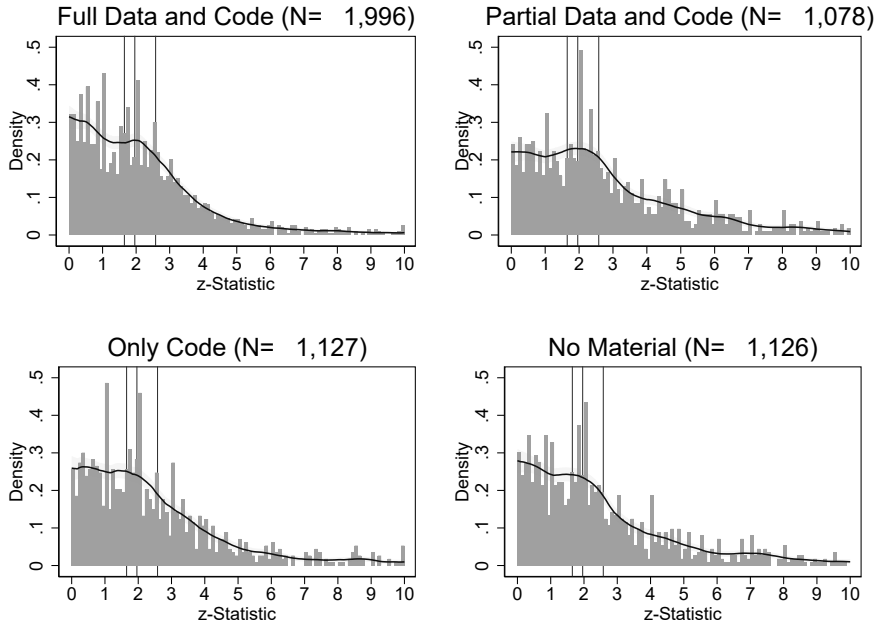
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ using a regression discontinuity design (RDD) by data type: admin, third-party survey, own-collected and other. We only consider those observations that rely solely on one type of data within each article. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A30: Estimated Density of z-Statistics using RCT by Data Type



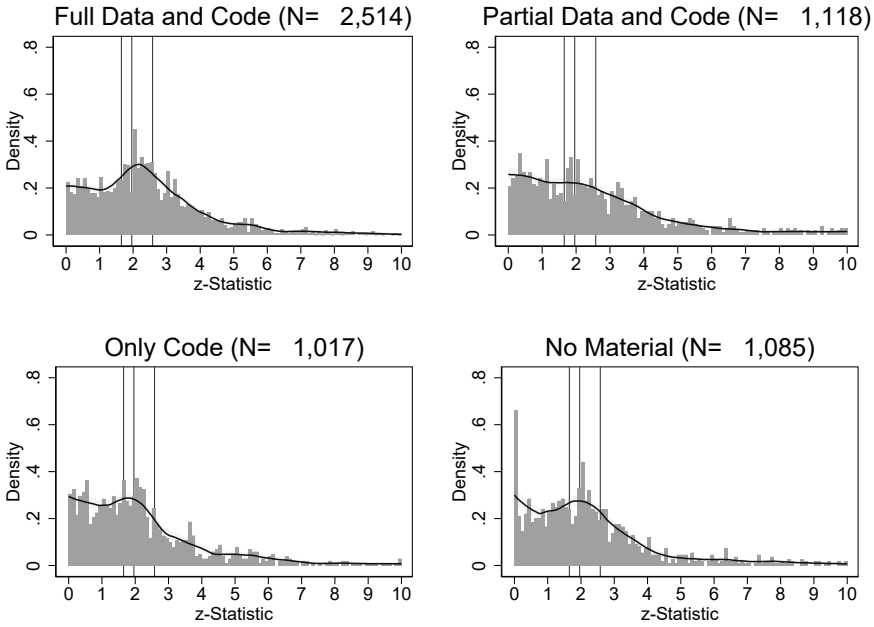
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ using a randomized control trials (RCT) by data type: admin, third- party survey, own-collected and other. We only consider those observations that rely solely on one type of data within each article. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates..

Figure A31: Estimated Density of z-Statistics using DID by Data and Code Availability



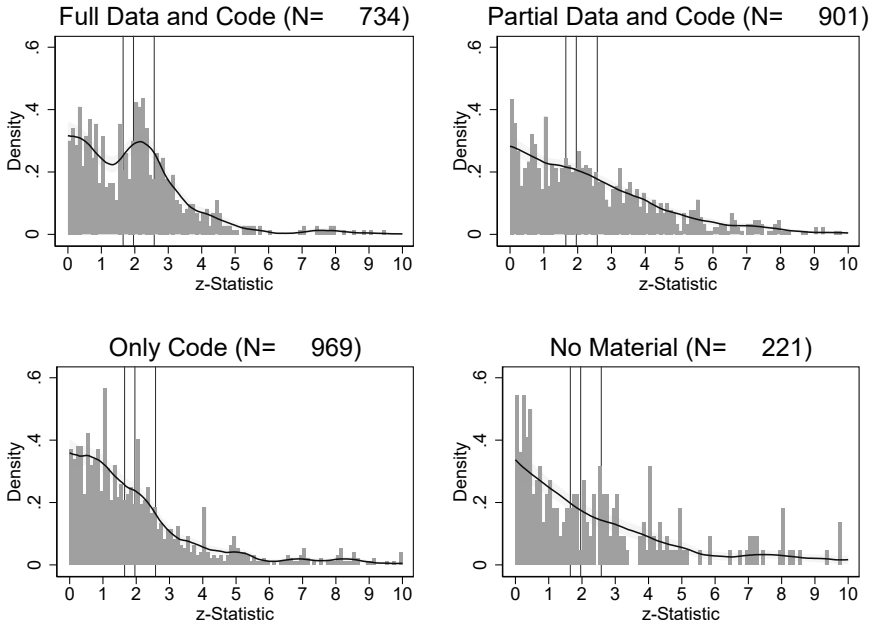
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ using a difference-in-differences (DID) approach by data and code availability: full data and code, partial data and code, only code and no material. We only consider those observations that rely solely on one type of data within each article. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A32: Estimated Density of z-Statistics using IV by Data and Code Availability



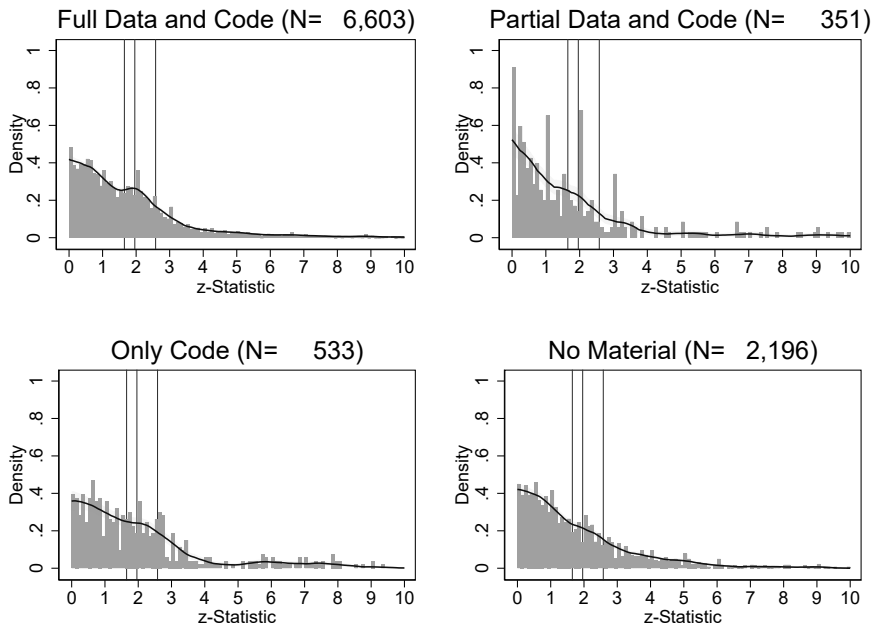
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ using a instrumental variables (IV) approach by data and code availability: full data and code, partial data and code, only code and no material. We only consider those observations that rely solely on one type of data within each article. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A33: Estimated Density of z-Statistics using RDD by Data and Code Availability



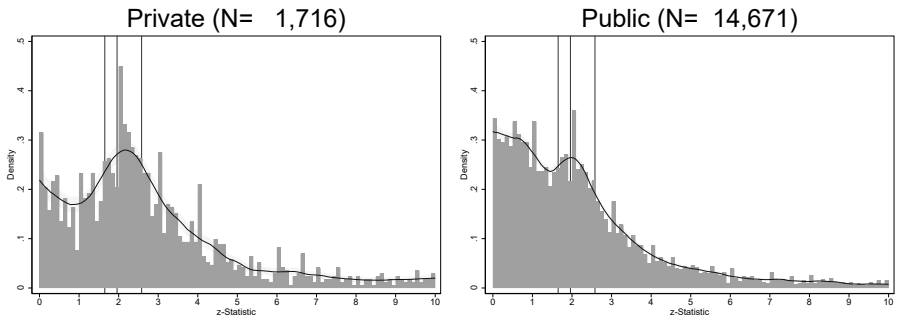
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$using a regression discontinuity design (RDD) by data and code availability: full data and code, partial data and code, only code and no material. We only consider those observations that rely solely on one type of data within each article. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A34: Estimated Density of z-Statistics using RCT by Data and Code Availability



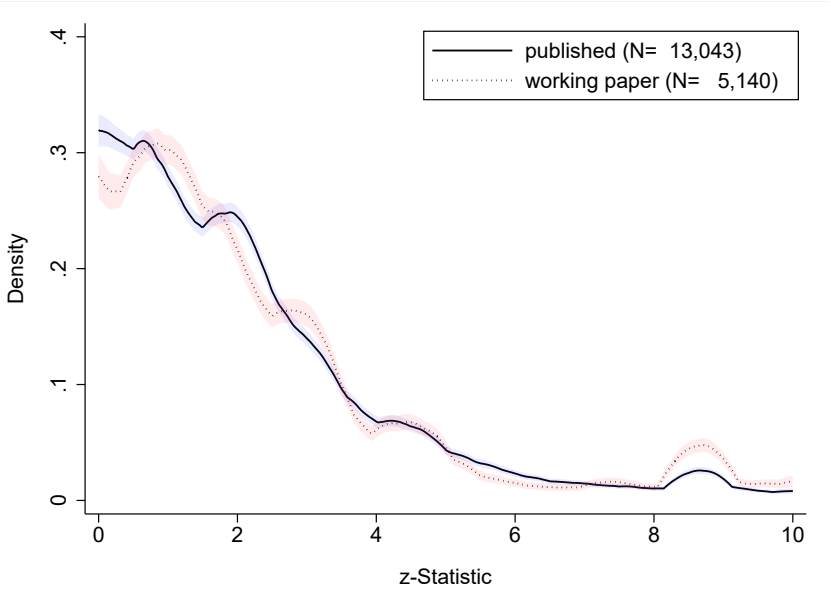
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ using a randomized control trials (RCT) by data and code availability: full data and code, partial data and code, only code and no material. We only consider those observations that rely solely on one type of data within each article. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A35: Estimated Density of z-Statistics: Private vs. Public Data



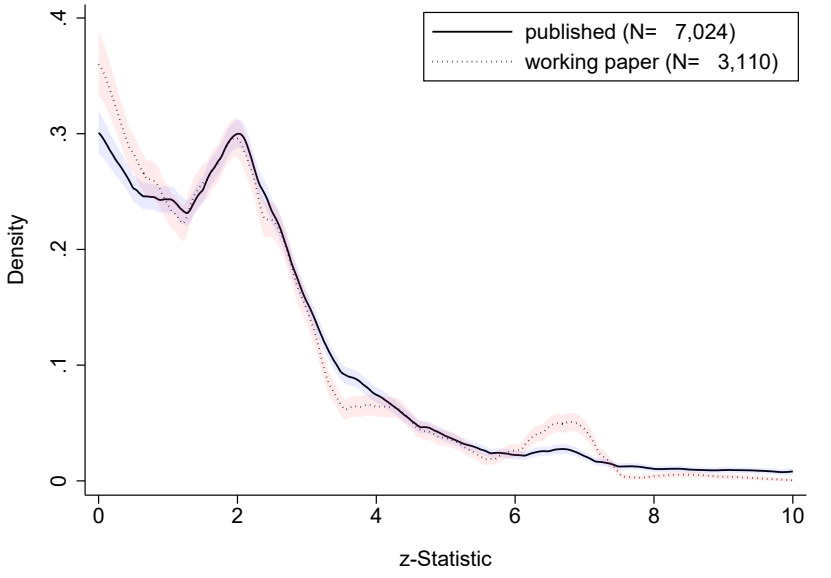
Notes: This figure displays 16,387 of test statistics for $z \in [0, 10]$. For clarity we only display test statistics that uniquely belong to the *private* or *public* type and ignore *own collection*. The sample is restricted to articles using only one type of data. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A36: Estimated Density by Publication Status and Data and Code Availability: Journals with Data-Policy



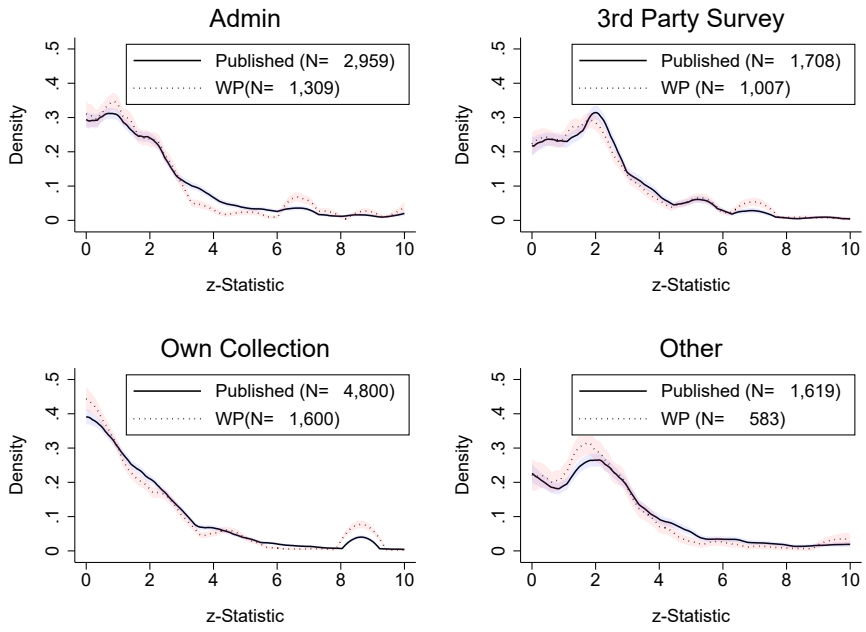
Notes: This sample is restricted to articles that are published in journals with a data-sharing policy. This figure displays two distributions. First, the solid line plots the z-statistics for those estimates that are *published* and second, the dotted line plots the z-statistics for those estimates that are published in a *working paper*. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A37: Estimated Density by Publication Status and Data and Code Availability: Journals without Data-Policy



Notes: This sample is restricted to articles that are published in journals without a data-sharing policy. This figure displays two distributions. First, the solid line plots the z-statistics for those estimates that are *published* and second, the dotted line plots the z-statistics for those estimates that are published in a *working paper*. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Figure A38: Estimated Density by Publication Status and Data Type



Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by data type: admin, third-party survey, own-collected and other. The solid line represent published z-statistics, while the dashed line represent those from working papers. The samples is accordingly restricted to estimates from published articles that had an associated working paper. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We also provide a kernel density with data-driven bandwidth. We do not weight our estimates.

Appendix Tables: NOT FOR PUBLICATION

Table A1: Summary Statistics: Data Availability and Journal Data-Sharing Policy

Full Sample					
	Full Data	Partial Data	Only Code	No Material	Total
Total Articles	302	123	106	576	1107
Articles in %	27.28	11.11	9.58	52.03	100
Total Tests	12579	3828	3954	18513	38874
Tests in %	32.36	9.85	10.17	47.62	100
Sample with Policy					
	Full Data	Partial Data	Only Code	No Material	Total
Total Articles	293	122	102	137	654
Articles in %	44.80	18.65	15.60	20.95	100
Total Tests	12366	3666	3848	4861	24741
Tests in %	49.98	14.82	15.55	19.65	100
Sample w/o Policy					
	Full Data	Partial Data	Only Code	No Material	Total
Total Articles	9	1	4	439	453
Articles in %	1.99	0.22	0.88	96.91	100
Total Tests	213	162	106	13652	14133
Tests in %	1.507	1.146	0.750	96.60	100
Pure Sample					
	Full Data	Partial Data	Only Code	No Material	Total
Total Articles	163	44	56	334	597
Articles in %	27.30	7.37	9.38	55.95	100
Total Tests	6981	1490	1770	10460	20701
Tests in %	33.72	7.198	8.550	50.53	100

Notes: The first part of this table provides an overview of the distribution of test statistics and total articles by data and code availability for the full sample. The second and third parts restrict the sample to articles published in a journal with and without a data-sharing policy, respectively. The bottom panel restrict the sample to articles relying solely on one data type.

Table A2: Summary Statistics: Data and Code Availability by Journal

Journal \ Provision of:	Full Data and Code	Partial Data and Code	Only Code	No Material
American Econ. J.: Applied Econ.	1,478	549	380	63
American Econ. J.: Econ. Policy	440	586	225	
American Econ. J.: Macroeconomics	18	24	12	
American Economic Review	3266	748	996	228
Econometrica	191	206		181
Economic Policy	21			59
Experimental Economics				79
Journal of Applied Econometrics	16	35		35
Journal of Development Economics	809	27	24	1958
Journal of Economic Growth	38			62
Journal of Financial Economics				569
Journal of Financial Intermediation				285
Journal of Human Resources	48		25	1624
Journal of International Economics			10	478
Journal of Labor Economics	387	219	146	362
Journal of Political Economy	1184	57	292	321
Journal of Public Economics				2605
Journal of Urban Economics	40			620
J. of the European Econ. Association	771	79	202	596
Review of Financial Studies				1618
The Economic Journal	960	272	424	973
The Journal of Finance	27	186	96	1775
The Quarterly Journal of Economics	673	109	3	3166
The Review of Economic Studies	1051	298	170	115
The Review of Economics and Statistics	1162	433	949	742
Total	12580	3828	3954	18514

Notes: This table provides an overview of test statistics and data and code availability by journal. It alphabetically presents our sample of Top 25 journals.

Table A3: Summary Statistics: Data Availability and Data Type

Pure Sample	Admin	Third-Party Survey	Own Collected	Other	Total
Total Articles	159	131	161	146	597
Articles in %	26.63	21.94	26.97	24.46	100
Total Tests	5726	3932	7754	3289	20701
Tests in %	27.66	18.99	37.46	15.89	100
Full Data and Code	Admin	Third-Party Survey	Own Collected	Other	Total
Total Articles	20	32	89	22	163
Articles in %	12.27	19.63	54.60	13.50	100
Total Tests	379	1165	4622	815	6981
Tests in %	5.429	16.69	66.21	11.67	100
Partial Data and Code	Admin	Third-Party Survey	Own Collected	Other	Total
Total Articles	25	10	2	7	44
Articles in %	56.82	22.73	4.545	15.91	100
Total Tests	754	283	248	205	1490
Tests in %	50.60	18.99	16.64	13.76	100
Only Code	Admin	Third-Party Survey	Own Collected	Other	Total
Total Articles	32	12	2	10	56
Articles in %	57.14	21.43	3.571	17.86	100
Total Tests	1231	272	35	232	1770
Tests in %	69.55	15.37	1.977	13.11	100
At Least Code	Admin	Third-Party Survey	Own Collected	Other	Total
Total Articles	77	54	93	39	263
Articles in %	29.28	20.53	35.36	14.83	100
Total Tests	2364	1720	4905	1252	10241
Tests in %	23.08	16.80	47.90	12.23	100
Nothing	Admin	Third-Party Survey	Own Collected	Other	Total
Total Articles	82	77	68	107	334
Articles in %	24.55	23.05	20.36	32.04	100
Total Tests	3362	2212	2849	2037	10460
Tests in %	32.14	21.15	27.24	19.47	100

Notes: The sample for all panels is restricted to test articles that rely solely on one data type within each article ("pure" sample). The first panel of this table provides an overview of the distribution of test statistics and articles by data and code availability and data type. The second panel restricts the sample to articles that provide full data and code. The third panel restricts the sample to articles providing partial data. The fourth panel restricts the sample to articles that provide only code. The fifth panel restricts the sample to articles that provide at least code. The last panel restricts the sample to articles that provide no data nor code.

Table A4: Summary Statistics: Data Type by Journal

Type of Data / Journal	Admin	3rd Party Survey	Own Collection	Other	Total
American Econ. J.: Applied Econ.	149	162	959	98	1,368
American Econ. J.: Econ. Policy	251	47	142	64	504
American Econ. J.: Macroeconomics		22			22
American Economic Review	779	590	1,460	526	3,355
Econometrica	126	86	106	5	323
Economic Policy		24		29	53
Experimental Economics			73		73
Journal of Applied Econometrics		35		16	51
Journal of Development Economics	51	153	1,321	120	1,645
Journal of Economic Growth		17		81	98
Journal of Financial Economics	40	9		269	318
Journal of Financial Intermediation		26		100	126
Journal of Human Resources	84	848	226	25	1,183
Journal of International Economics	148	206		48	402
Journal of Labor Economics	355	293	171		819
Journal of Political Economy	149	43	552	114	858
Journal of Public Economics	815	73	374	54	1,316
Journal of Urban Economics	284	14			298
J. of the European Econ. Association	197	206	328	150	881
Review of Financial Studies	63	46	62	521	692
The Economic Journal	527	362	284	118	1,291
The Journal of Finance	37		352	543	932
The Quarterly Journal of Economics	1,021	53	590	95	1,759
The Review of Economic Studies	134	61	370	96	661
The Review of Economics and Statistics	516	556	384	217	1,673
Total	5,726	3,932	7,754	3,289	20,701

Notes: This table provides an overview of test statistics by data type and journal. It alphabetically presents our sample of Top 25 journals. We only consider those estimates that rely solely on one type of data ("Pure" sample).

Table A5: Caliper Tests 10% Threshold: Data-Sharing

	Significant at 10% Level					
	(1)	(2)	(3)	(4)	(5)	(6)
Data & Code Provided (Author)	-0.036*	-0.016	-0.027	-0.015	-0.020	-0.006
	(0.020)	(0.019)	(0.020)	(0.020)	(0.023)	(0.024)
Data & Code Provided (Journal)	0.010	0.027	0.007	0.016	0.003	0.009
	(0.016)	(0.017)	(0.018)	(0.019)	(0.022)	(0.024)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	9,224	9,223	6,110	6,110	3,597	3,596
Threshold	1.65	1.65	1.65	1.65	1.65	1.65
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 10% level and zero otherwise. The primary independent variables take a value one if the test statistic is drawn from an article with data and code provided (on the author's homepage or journal's website) and zero otherwise.

Table A6: Caliper Tests 1% Threshold: Data-Sharing

	Significant at 1% Level					
	(1)	(2)	(3)	(4)	(5)	(6)
Data & Code Provided (Author)	0.022	0.025	-0.002	-0.000	0.004	0.000
	(0.022)	(0.022)	(0.024)	(0.024)	(0.031)	(0.030)
Data & Code Provided (Journal)	-0.009	0.001	-0.014	-0.008	-0.015	-0.006
	(0.019)	(0.019)	(0.021)	(0.021)	(0.026)	(0.026)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	7,185	7,181	4,949	4,946	2,854	2,851
Threshold	2.58	2.58	2.58	2.58	2.58	2.58
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 1% level and zero otherwise. The primary independent variables take a value one if the test statistic is drawn from an article with data and code provided (on the author's homepage or journal's website) and zero otherwise.

Table A7: Control Variables for the Caliper Tests 5% Threshold: Data-Sharing

	(1)	(2)	(3)	(4)	(5)	(6)
Data & Code Provided (Author)	-0.038*	-0.020	-0.042*	-0.031	-0.026	-0.021
	(0.020)	(0.021)	(0.023)	(0.023)	(0.028)	(0.029)
Data & Code Provided (Journal)	0.008	0.019	0.006	0.016	0.010	0.020
	(0.019)	(0.019)	(0.020)	(0.020)	(0.024)	(0.024)
Solo-Authored		-0.043*		-0.040		-0.044
		(0.025)		(0.027)		(0.034)
Experience		-0.001		0.001		0.007
		(0.003)		(0.004)		(0.004)
Experience^2		0.002		-0.005		-0.025*
		(0.010)		(0.012)		(0.013)
Editor Present		-0.048***		-0.031		-0.044*
		(0.018)		(0.020)		(0.024)
Top Institution		-0.034		-0.012		-0.036
		(0.031)		(0.035)		(0.041)
PhD Top Institution		-0.021		-0.025		0.002
		(0.030)		(0.033)		(0.039)
Share Female Authors		-0.012		-0.019		-0.009
		(0.027)		(0.030)		(0.037)
Identification: IV		0.015		-0.001		-0.019
		(0.019)		(0.021)		(0.026)
Identification: RCT		-0.034*		-0.033		-0.055**
		(0.020)		(0.022)		(0.027)
Identification: RDD		-0.019		-0.018		-0.032
		(0.023)		(0.025)		(0.029)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Observations	9,334	9,332	6,829	6,828	4,060	4,059
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: Appendix Table 4 with controls exposed. Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variables take a value one if the test statistic is drawn from an article with data and code provided (on the author's homepage or journal's website) and zero otherwise.

Table A8: Caliper Tests 10% Threshold: Data-Sharing and Data-Sharing Policy

	(1)	(2)	(3)	(4)	(5)	(6)
First Stage			DV: Data and Code Provided			
Data & Code Policy	0.839***	0.846***	0.889	0.926***	0.940***	1.028***
	(0.156)	(0.168)	(.)	(0.177)	(0.165)	(0.204)
Second Stage			DV: Significant at 10% Level			
Data & Code Provided	0.139	0.177	0.179	0.307	0.400	0.519
(Journal)	(0.268)	(0.332)	(0.279)	(0.317)	(0.328)	(0.370)
Reduced Form			DV: Significant at 10% Level			
Data & Code Policy	0.030	0.031	0.041	0.058	0.095	0.107
	(0.056)	(0.060)	(0.062)	(0.062)	(0.075)	(0.075)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	3,222	3,222	2,160	2,160	1,288	1,287
Threshold	1.65	1.65	1.65	1.65	1.65	1.65
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: **All Panels:** Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. **First Stage:** Requiring data and code to be posted increases the provision of data and code. The dependent variable takes a value one if the test statistic is drawn from an article with data and code provided on the journal's website and zero otherwise. The primary independent variable takes a value one if the article was published in a journal up to 5 years after the journal implemented a data and code availability requirement and zero up to 5 years before the journal implements a policy. **Second Stage:** Provision of data and code, instrumented by journal policy, does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 10% level and zero otherwise. The primary independent variable takes a value one if the article was predicted to provide data and code on the journal's website using journal policy as an instrument. **Reduced Form:** Requiring data and code to be posted does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 10% level and zero otherwise. The primary independent variable takes a value one if the article is published in a journal with a data and code policy.

Table A9: Caliper Tests 1% Threshold: Data-Sharing and Data-Sharing Policy

	(1)	(2)	(3)	(4)	(5)	(6)
First Stage			DV: Data and Code Provided			
Data & Code Policy	0.623***	0.574***	0.623***	0.560***	0.667***	0.490***
	(0.184)	(0.161)	(0.187)	(0.161)	(0.179)	(0.151)
Second Stage			DV: Significant at 1% Level			
Data & Code Provided	0.745	0.815	0.903	1.138*	0.608	0.977
(Journal)	(0.509)	(0.565)	(0.636)	(0.692)	(0.580)	(0.679)
Reduced Form			DV: Significant at 1% Level			
Data & Code Policy	0.116*	0.110*	0.135*	0.154**	0.113	0.152
	(0.066)	(0.064)	(0.076)	(0.074)	(0.096)	(0.094)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	2,390	2,389	1,643	1,642	959	958
Threshold	2.58	2.58	2.58	2.58	2.58	2.58
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: **All Panels:** Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. **First Stage:** Requiring data and code to be posted increases the provision of data and code. The dependent variable takes a value one if the test statistic is drawn from an article with data and code provided on the journal's website and zero otherwise. The primary independent variable takes a value one if the article was published in a journal up to 5 years after the journal implemented a data and code availability requirement and zero up to 5 years before the journal implements a policy. **Second Stage:** Provision of data and code, instrumented by journal policy, does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 1% level and zero otherwise. The primary independent variable takes a value one if the article was predicted to provide data and code on the journal's website using journal policy as an instrument. **Reduced Form:** Requiring data and code to be posted does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 1% level and zero otherwise. The primary independent variable takes a value one if the article is published in a journal with a data and code policy.

Table A10: Instrumental Variable Approach (OLS)

	(1)	(2)	(3)	(4)	(5)	(6)
First Stage			DV: Data and Code Provided			
Data & Code Policy	0.462***	0.402**	0.407**	0.340**	0.424**	0.363**
	(0.171)	(0.155)	(0.172)	(0.153)	(0.174)	(0.154)
Second Stage			DV: Significant at 5% Level			
Data & Code Provided	-0.099	-0.198	-0.032	-0.095	0.114	0.104
	(0.152)	(0.176)	(0.176)	(0.192)	(0.199)	(0.213)
Reduced Form			DV: Significant at 5% Level			
Data & Code Policy	-0.046	-0.079	-0.013	-0.032	0.048	0.038
	(0.066)	(0.064)	(0.072)	(0.066)	(0.084)	(0.076)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	3,177	3,177	2,282	2,282	1,362	1,362
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20
First Stage F	144.87	105.79	81.35	56.00	53.18	38.70

Notes: Each observation is a test statistic. We rely on OLS models and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. **First Stage:** Requiring data and code to be posted increases the provision of data and code. The dependent variable takes a value one if the test statistic is drawn from an article with data and code provided on the journal's website and zero otherwise. The primary independent variable takes a value one if the article was published in a journal up to 5 years after the journal implemented a data and code availability requirement and zero up to 5 years before the journal implements a policy. **Second Stage:** Provision of data and code, instrumented by journal policy, does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variable takes a value one if the article was predicted to provide data and code on the journal's website using journal policy as an instrument. **Reduced Form:** Requiring data and code to be posted does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variable takes a value one if the article is published in a journal with a data and code policy.

Table A11: Article Weights: Caliper Tests 5% Threshold: Data-Sharing

	Significant at 5% Level (Article Weights)					
	(1)	(2)	(3)	(4)	(5)	(6)
Data Or Code Provided (Author)	0.006	0.018	-0.004	0.001	-0.021	-0.024
	(0.028)	(0.028)	(0.030)	(0.030)	(0.040)	(0.039)
Data Or Code Provided (Journal)	-0.023	-0.029	-0.019	-0.026	-0.010	-0.015
	(0.030)	(0.030)	(0.032)	(0.032)	(0.039)	(0.039)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	9,334	9,332	6,829	6,828	4,060	4,059
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: This table replicates Table 4 after applying article weights. Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variables take a value one if the test statistic is drawn from an article with data and code provided (on the author's homepage or journal's website) and zero otherwise.

Table A12: De-Rounding: Caliper Tests 5% Threshold: Data-Sharing

	Significant at 5% Level (Derounded)					
	(1)	(2)	(3)	(4)	(5)	(6)
Data Or Code Provided (Author)	-0.024	-0.015	-0.028	-0.026	-0.016	-0.024
	(0.019)	(0.020)	(0.022)	(0.023)	(0.025)	(0.027)
Data Or Code Provided (Journal)	-0.028	-0.065**	-0.044	-0.076**	-0.032	-0.064*
	(0.026)	(0.030)	(0.028)	(0.033)	(0.031)	(0.037)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	9,329	9,329	6,825	6,825	4,057	4,057
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: This table replicates Table 4 after applying a de-rounding method as in Brodeur et al. (2016). Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variables take a value one if the test statistic is drawn from an article with data and code provided (on the author's homepage or journal's website) and zero otherwise.

Table A13: Article Weights: Caliper Tests 5% Threshold: Data-Sharing and Data-Sharing Policy

	(1)	(2)	(3)	(4)	(5)	(6)
First Stage			DV: Data and Code Provided			
Data & Code Policy	0.556***	0.535***	0.480***	0.463***	0.479***	0.479***
	(0.175)	(0.164)	(0.175)	(0.155)	(0.179)	(0.165)
Second Stage			DV: Significant at 5% Level			
Data & Code Provided	-1.143**	-1.266*	-1.209*	-1.203	-0.646	-0.556
(Journal)	(0.565)	(0.648)	(0.622)	(0.755)	(0.790)	(0.870)
Reduced Form			DV: Significant at 5% Level			
Data & Code Policy	-0.143*	-0.134*	-0.127	-0.104	-0.066	-0.050
	(0.075)	(0.074)	(0.081)	(0.077)	(0.093)	(0.089)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	3,177	3,177	2,282	2,282	1,361	1,361
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: This table replicates Table 5 after applying article weights. Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. **First Stage:** Requiring data and code to be posted increases the provision of data and code. The dependent variable takes a value one if the test statistic is drawn from an article with data and code provided on the journal's website and zero otherwise. The primary independent variable takes a value one if the article was published in a journal up to 5 years after the journal implemented a data and code availability requirement and zero up to 5 years before the journal implements a policy. **Second Stage:** Provision of data and code, instrumented by journal policy, does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variable takes a value one if the article was predicted to provide data and code on the journal's website using journal policy as an instrument. **Reduced Form:** Requiring data and code to be posted does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variable takes a value one if the article is published in a journal with a data and code policy.

Table A14: De-Rounding: Caliper Tests 5% Threshold: Data-Sharing and Data-Sharing Policy

	(1)	(2)	(3)	(4)	(5)	(6)
First Stage			DV: Data and Code Provided			
Data & Code Policy	0.797***	0.788***	0.730***	0.675***	0.725***	0.697***
	(0.168)	(0.175)	(0.173)	(0.162)	(0.174)	(0.172)
Second Stage			DV: Significant at 5% Level			
Data & Code Provided	-0.378	-0.755*	-0.212	-0.535	0.102	-0.100
(Journal)	(0.400)	(0.450)	(0.446)	(0.504)	(0.488)	(0.537)
Reduced Form			DV: Significant at 5% Level			
Data & Code Policy	-0.069	-0.122*	-0.033	-0.070	0.017	-0.014
	(0.066)	(0.064)	(0.071)	(0.067)	(0.080)	(0.075)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	3,174	3,174	2,279	2,279	1,359	1,359
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: This table replicates Table 5 after applying a de-rounding method as in Brodeur et al. (2016). Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. **First Stage:** Requiring data and code to be posted increases the provision of data and code. The dependent variable takes a value one if the test statistic is drawn from an article with data and code provided on the journal's website and zero otherwise. The primary independent variable takes a value one if the article was published in a journal up to 5 years after the journal implemented a data and code availability requirement and zero up to 5 years before the journal implements a policy. **Second Stage:** Provision of data and code, instrumented by journal policy, does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variable takes a value one if the article was predicted to provide data and code on the journal's website using journal policy as an instrument. **Reduced Form:** Requiring data and code to be posted does not affect statistical significance. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variable takes a value one if the article is published in a journal with a data and code policy.

Table A15: Results from Application of Elliott et al. (2022) at the 10% and 1% Significance Thresholds

10% Threshold	Total	Obs	Bin	Disc.	LCM	CS1	CS2B
Full Sample							
Full Data and Code	7368	164	0.002	0.120	0.006	0.002	0.000
Other	15928	328	0.391	0.014	0.000	0.000	0.000
Pre-Journal Policy							
Full Data and Code	103	0	.	.	.	.	.
Other	1657	45	0.036	0.664	0.794	0.140	0.000
Post-Journal Policy							
Full Data and Code	2428	53	0.006	0.053	0.179	0.002	0.000
Other	3682	63	0.401	0.000	0.418	0.015	0.000
(Only Journals That Switch)							
Full Data and Code	2531	53	0.006	0.042	0.185	0.002	0.000
Other	5339	108	0.074	0.001	0.140	0.006	0.000
Data Type							
Admin	3681	72	0.638	0.072	0.253	0.001	0.001
Own	4106	88	0.002	0.024	0.071	0.025	0.006
Survey	2343	54	0.752	0.060	0.444	0.000	0.000
Other	3319	58	0.448	0.077	0.133	0.601	0.202
1% Threshold	Total	Obs	Bin	Disc.	LCM	CS1	CS2B
Full Sample							
Full Data and Code	7368	894	1.000	0.194	0.006	0.002	0.000
Other	15928	1681	1.000	0.601	0.000	0.000	0.000
Pre-Journal Policy							
Full Data and Code	103	19	.	.	.	.	.
Other	1657	181	0.617	0.787	0.794	0.140	0.000
Post-Journal Policy							
Full Data and Code	2428	292	0.955	0.683	0.179	0.002	0.000
Other	3682	374	1.000	0.783	0.418	0.015	0.000
(Only Journals That Switch)							
Full Data and Code	2531	311	0.979	0.552	0.185	0.002	0.000
Other	5339	555	0.999	0.815	0.14	0.006	0.000
Data Type							
Admin	3681	368	0.973	0.717	0.253	0.001	0.001
Own	4106	465	1.000	0.938	0.071	0.025	0.006
Survey	2343	270	0.961	0.884	0.444	0.000	0.000
Other	3319	392	0.999	0.822	0.133	0.601	0.202

Notes: This table provides the result from the battery of tests proposed in Elliott et al. (2022) for the 10% and 1% Thresholds for both data-sharing and data-type analysis.

Table A16: Prediction of Provision of Full Data and at Least Code: Journals with Data-Sharing Policy

Provision of ...	Full Data and Code			At least Code		
	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
Third-Party Survey	0.428***	0.380***	0.379***	0.047	0.044	0.290
	(0.088)	(0.081)	(0.071)	(0.095)	(0.060)	(0.332)
Own-Collected	0.585***	0.421***	0.455***	-0.014	-0.059	0.098
	(0.073)	(0.085)	(0.079)	(0.094)	(0.075)	(0.374)
Other	0.357***	0.191**	0.212**	-0.037	-0.122	-0.283
	(0.128)	(0.088)	(0.089)	(0.111)	(0.090)	(0.390)
Controls						
DID		–0.126*	–0.161**		-0.029	-0.320
		(0.068)	(0.065)		(0.058)	(0.289)
IV		–0.113*	–0.131**		0.039	0.164
		(0.061)	(0.059)		(0.059)	(0.302)
RDD		–0.199**	–0.267***		0.107	0.217
		(0.100)	(0.087)		(0.087)	(0.439)
Top 5		0.228***	-0.059		0.300***	1.044**
		(0.052)	(0.096)		(0.056)	(0.472)
Experience		0.013	0.017*		0.000	0.011
		(0.011)	(0.010)		(0.011)	(0.052)
Experience^2		-0.026	-0.031		0.003	0.031
		(0.031)	(0.027)		(0.030)	(0.151)
Top Institution		-0.118	-0.077		-0.083	-0.033
		(0.106)	(0.090)		(0.087)	(0.384)
PhD Top Institution		-0.123	–0.164*		-0.020	-0.219
		(0.107)	(0.091)		(0.085)	(0.342)
Other Controls						
Year FE	Y	Y	Y	Y	Y	Y
Solo Authored		Y	Y		Y	Y
Share Female Authors		Y	Y		Y	Y
Editor		Y	Y		Y	Y
Field			Y			Y
Observations	13,220	12,950	12,788	13,220	12,677	12,151

Notes: We rely on probit models and present the average marginal effects (equation (3)). We restrict the sample to journals with a data-sharing policy. The dependent variable in column (1)-(3) is a dummy for whether full data and code can be accessed, while the dependent variable is a dummy for whether at least code is available on webpages of the journals for column (4)-(6). The omitted category is *admin*. The omitted category for the methods is RCT. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table A17: Results from Application of Elliott et al. (2022): Data Type

Sample	Bin.	Disc.	CS1	CS2B	LCM	Obs	Total
Admin	0.033	0.717	0.001	0.001	0.253	171	3681
Own Collected	0.000	0.938	0.025	0.006	0.071	252	4106
3rd Party Survey	0.000	0.884	0.000	0.000	0.444	123	2343
Other	0.000	0.822	0.601	0.202	0.133	163	3319

Notes: This table provides the result from the battery of tests proposed in Elliott et al. (2022) for the 5% threshold for data type. We restrict the sample to articles relying solely on one data type.

Table A18: Article Weights: Caliper Tests 5% Threshold: Data Type

	Significant at 5% Level (Article Weights)					
	(1)	(2)	(3)	(4)	(5)	(6)
Own-Collected	-0.032	-0.011	-0.011	0.012	0.031	0.040
	(0.031)	(0.049)	(0.034)	(0.050)	(0.046)	(0.061)
Third-Party Survey	-0.018	-0.028	-0.011	-0.012	0.090*	0.075
	(0.033)	(0.033)	(0.036)	(0.036)	(0.052)	(0.053)
Other	-0.019	-0.036	-0.037	-0.057	-0.002	-0.009
	(0.041)	(0.038)	(0.045)	(0.043)	(0.055)	(0.055)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Data and Code Provided		✓		✓		✓
Additional Controls		✓		✓		✓
Observations	4,814	4,814	3,517	3,517	2,089	2,089
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: This table replicates Table 9 after applying article weights. Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variables are dummy variables for each data type.

Table A19: De-Rounding: Caliper Tests 5% Threshold: Data Type

	Significant at 5% Level (Derounded)					
	(1)	(2)	(3)	(4)	(5)	(6)
Own-Collected	-0.013	0.048	0.002	0.075*	0.007	0.038
	(0.024)	(0.036)	(0.028)	(0.039)	(0.032)	(0.049)
Third-Party Survey	-0.013	-0.006	-0.031	-0.006	-0.033	-0.036
	(0.032)	(0.028)	(0.033)	(0.030)	(0.039)	(0.039)
Other	-0.046	-0.075**	-0.035	-0.060*	-0.034	-0.079*
	(0.033)	(0.032)	(0.035)	(0.035)	(0.042)	(0.043)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Data and Code Provided		✓		✓		✓
Additional Controls		✓		✓		✓
Observations	4,811	4,811	3,514	3,514	2,087	2,087
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: This table replicates Table 9 after applying a de-rounding method as in Brodeur et al. (2016). Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variables are dummy variables for each data type.

Table A20: Caliper Tests 10% Threshold: Data Type

	Significant at 10% Level					
	(1)	(2)	(3)	(4)	(5)	(6)
Own-Collected	0.013	0.015	0.029	0.017	0.004	-0.018
	(0.024)	(0.037)	(0.027)	(0.045)	(0.034)	(0.050)
Third-Party Survey	0.045	0.022	0.050	0.025	0.017	-0.029
	(0.028)	(0.028)	(0.032)	(0.033)	(0.040)	(0.042)
Other	0.088***	0.072**	0.088**	0.068*	0.041	0.015
	(0.032)	(0.032)	(0.037)	(0.037)	(0.047)	(0.050)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Data and Code Provided		✓		✓		✓
Additional Controls		✓		✓		✓
Observations	4,832	4,832	3,198	3,198	1,853	1,852
Threshold	1.65	1.65	1.65	1.65	1.65	1.65
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 10% level and zero otherwise. The primary independent variables are dummy variables for each data type.

Table A21: Caliper Tests 1% Threshold: Data Type

	Significant at 1% Level					
	(1)	(2)	(3)	(4)	(5)	(6)
Own-Collected	0.017	0.006	0.025	0.003	0.039	0.019
	(0.028)	(0.036)	(0.029)	(0.042)	(0.039)	(0.057)
Third-Party Survey	0.047	0.055	0.036	0.056	0.001	0.016
	(0.034)	(0.036)	(0.035)	(0.038)	(0.046)	(0.049)
Other	0.079**	0.091**	0.086**	0.091**	0.125***	0.132**
	(0.034)	(0.036)	(0.038)	(0.040)	(0.048)	(0.051)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Data and Code Provided		✓		✓		✓
Additional Controls		✓		✓		✓
Observations	3,680	3,679	2,549	2,548	1,477	1,476
Threshold	2.58	2.58	2.58	2.58	2.58	2.58
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 1% level and zero otherwise. The primary independent variables are dummy variables for each data type.

Table A22: Relative Publication Probabilities by Data Type

Sample	u	t	df	[0,1.1645]	(1.645,1.960]	(1.960,2.576]
Admin	0.001	0.001	1.016	0.435	0.876	1.237
	0.000	0.001	0.021	0.026	0.058	0.063
Own-Collected	0.016	0.012	1.649	0.495	0.856	1.199
	0.002	0.002	0.044	0.030	0.055	0.065
Third-Party Survey	0.006	0.005	2.048	0.288	0.711	0.975
	0.001	0.001	0.074	0.019	0.054	0.062
Other	0.001	0.003	1.425	0.202	0.593	0.944
	0.000	0.001	0.036	0.009	0.039	0.049

Notes: The table presents the results of applying the publication bias model presented in Andrews and Kasy (2019) to data type. The model assumes that the underlying effect sizes follow a generalized t distribution. We report the model's estimated location parameter, scale parameter, and degrees of freedom in the first three columns. In the fourth column, 0.435 represents the relative probability that a test statistic in the [0,1.1645] interval is 43.5% as likely to be published as a test statistic greater than 2.576 (the reference interval).

Table A23: Public Versus Private Data: Caliper Test 5% Threshold

	(1)	(2)	(3)	(4)	(5)	(6)
Public Data	-0.042	-0.049	-0.050	-0.043	0.006	-0.002
	(0.039)	(0.038)	(0.042)	(0.041)	(0.058)	(0.061)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	4,175	4,175	3,042	3,042	1,800	1,800
Threshold	1.96	1.96	1.96	1.96	1.96	1.96
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 5% level and zero otherwise. The primary independent variable is a dummy variable that takes the value one if the dataset is categorized as 'public' and zero if 'private'.

Table A24: Public Versus Private Data: Caliper Test 10% Threshold

	Significant at 10% Level					
	(1)	(2)	(3)	(4)	(5)	(6)
Public Data	-0.009	-0.021	0.025	0.012	-0.003	-0.028
	(0.050)	(0.048)	(0.061)	(0.058)	(0.076)	(0.075)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	4,082	4,082	2,715	2,715	1,581	1,580
Threshold	1.65	1.65	1.65	1.65	1.65	1.65
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 10% level and zero otherwise. The primary independent variable is a dummy variable that takes the value one if the dataset is categorized as 'public' and zero if 'private'.

Table A25: Public Versus Private: Caliper Test 1% Threshold

	Significant at 1% Level					
	(1)	(2)	(3)	(4)	(5)	(6)
Public Data	-0.006	-0.015	-0.027	-0.026	-0.049	-0.044
	(0.045)	(0.046)	(0.047)	(0.050)	(0.049)	(0.052)
Year FE	✓	✓	✓	✓	✓	✓
Journal FE	✓	✓	✓	✓	✓	✓
Additional Controls		✓		✓		✓
Observations	3,280	3,279	2,253	2,252	1,282	1,281
Threshold	2.58	2.58	2.58	2.58	2.58	2.58
Window	0.50	0.50	0.35	0.35	0.20	0.20

Notes: Each observation is a test statistic. We rely on probit models and present the average marginal effects and associated standard errors clustered at the journal article-level. Observations are unweighted. Additional controls include identification strategy fixed effects, dummy variables for how results are reported, a dummy for solo-authorship, average years since PhD, average years since PhD squared, share of authors graduated from top PhD institution, share of authors at top institution, 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. The dependent variable takes a value one if the test statistic is significant at the 1% level and zero otherwise. The primary independent variable is a dummy variable that takes the value one if the dataset is categorized as 'public' and zero if 'private'.

Table A26: Working Paper Available?

	(1)	(2)	(3)	(4)
Method of Data Collection: (omitted admin)				
third party survey	0.082	0.064	0.069	0.035
	(0.074)	(0.075)	(0.077)	(0.077)
own collection	0.055	0.045	0.033	0.038
	(0.064)	(0.065)	(0.066)	(0.065)
other	-0.089	-0.114*	-0.051	-0.031
	(0.066)	(0.067)	(0.075)	(0.073)
Other Controls				
Reporting Method		Y	Y	Y
Solo Authored		Y	Y	Y
Share Female Authors		Y	Y	Y
Editor		Y	Y	Y
Field FE			Y	
Journal FE				Y
Articles	404	404	404	397

Notes: We rely on probit models and present the average marginal effects (equation (4)). The dependent variable is a dummy that takes a value of one if a published article has a public working paper. No article weights applied.

Table A27: Working Paper vs Published Version: Caliper Test 5% Threshold

	(1) ALL	(2) Admin	(3) 3rd Party Survey	(4) Own Collected	(5) Other
Published Version	-0.015	-0.027	0.027	-0.021	-0.089
	(0.018)	(0.038)	(0.031)	(0.028)	(0.058)
Test Statistics	3,818	622	552	833	138
Articles	332	41	29	51	14
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]

Notes: This table reports estimates from a linear probability regression with article fixed effects. The dependent variable is a dummy that takes a value one if a given test statistic is significant at the 5% level (i.e. equal to 1.96). The independent variable of interest is a dummy that takes the value of one if a given test statistic is from the published version of an article. The sample is accordingly restricted to estimates from published articles that had an associated working paper. We apply no weights. The analysis is based on the "pure" sample (i.e., solely one data type per article).