Methods Matter: P-Hacking and Causal Inference in Economics and Finance*

Abel Brodeur Nikolai Cook Anthony Heyes

March 2019

Abstract

The economics 'credibility revolution' has promoted the identification of causal relationships using difference-in-differences (DID), instrumental variables (IV), randomized control trials (RCT) and regression discontinuity design (RDD) methods. The extent to which a reader should trust claims about the statistical significance of results proves very sensitive to method. Applying multiple methods to over 12,000 hypothesis tests reported in 25 top economics journals in 2015, we show that selective publication and p-hacking is a substantial problem in research employing DID and (in particular) IV. RCT and RDD are much less problematic. About 15% of claims of marginally significant results in IV papers are misleading.

KEYWORDS: Research methods - causal inference - p-curves - p-hacking - publication bias

JEL CODES: A11, B41, C13, C44.

*Brodeur: Department of Economics, University of Ottawa, 120 University, Social Sciences Building, Ottawa, Ontario K1N 6N5, Canada. E-mail: abrodeur@uottawa.ca. Cook: Department of Economics, University of Ottawa, 120 University, Social Sciences Building, Ottawa, Ontario K1N 6N5, Canada. E-mail: ncook@uottawa.ca. Heyes: Department of Economics, University of Ottawa, 120 University Private, Ottawa, Ontario, Canada, K1N 6N5 and University of Sussex, E-mail: anthony.heyes@uottawa.ca. We are grateful to Andrew Foster, Jason Garred, Fernando Hoces de la Guardia, Jon de Quidt, Matt Webb and seminar participants at the BITSS annual meeting, Carleton University and the University of Ottawa for useful remarks and encouragement. We thank Richard Beard and Jessica Krueger for research assistance. Errors are ours.

In recent years, empirical economics has experienced a "credibility revolution" with a shift in focus to causal inference (Angrist and Pischke, 2010). As a result, experimental and quasi-experimental methods such as difference-in-differences (DID), instrumental variables (IV), randomized control trials (RCT), and regression discontinuity design (RDD) have become the norm in applied microeconomics (Panhans and Singleton (2017)). The rapid growth in the use of these methods is due to many reasons, one of which being the increasing demand for evidence-based policy. Experimental and quasi-experimental methods allow researchers to conduct impact evaluation and are, arguably, better suited at uncovering causal effects than naive correlation analysis.

We hold these methods to a higher standard by comparing them *to each other*. Our interest is in the extent to which claims made about statistical significance in papers using each of these methods are reliable. Evidence of publication bias and p-hacking in economics and other disciplines is by now voluminous (for examples Casey et al. (2012); Gerber and Malhotra (2008a); Gerber and Malhotra (2008b); Havránek (2015); Henry (2009); Ioannidis (2005); Ioannidis et al. (2017); Leamer (1983); Leamer and Leonard (1983); McCloskey (1985); Ridley et al. (2007); Simmons et al. (2011); Stanley (2008); Vivalt (2019)). Publication bias, whereby the statistical significance of a result determines the likelihood it is published, is likely a reflection of the peer review process. P-hacking refers to a variety of practices that a researcher might use to generate 'better' p-values, perhaps (but not necessarily) in response to the difficulty of publishing insignificant results (Abadie (2018); Blanco-Perez and Brodeur (2017); Doucouliagos and Stanley (2013); Franco et al. (2014); Furukawa (2017); Rosenthal (1979); Stanley (2005)). Such practices might include continuing to collecting data until a significance threshold is met, re-selecting covariates, or imposing sample restrictions in order to move a test statistic across a significance threshold.

The aim in this paper is to address systematically the questions: (1) How reliable are claims made about the statistical significance of causal relationships published in highly-ranked economics journals? (2) To what extent does the answer to that question depend upon the method of inference used? These questions should be of interest to policymakers who use empirical evidence to inform decisions and policies, as specification searching and

data mining will create literatures with a high percentage of false positives.

In order to answer these questions, we harvest the universe of hypothesis tests (13,440 in total) reported in papers using these methods in 25 top economics journals during 2015. We show that, from this perspective on credibility, methods matter. As a whole, the distribution of published test statistics has a two-humped camel shape, with 'missing' tests just before the conventional significance thresholds, i.e., $z = 1.65$, and a 'surplus' just after (Brodeur et al., 2016). Strikingly, this pattern is much more prevalent in research employing DID and (in particular) IV than in articles using RCT or RDD.

We employ three approaches in our analysis to document the extent of publication bias and p-hacking by methods.

First, we use probit regressions to study the extent to which the likelihood that a test delivers a significant result is sensitive to the method employed. Using RCT as a baseline, we find that journal articles using DID and IV are about 15% more likely to report test statistics that are statistically significant at the 5 percent level. In contrast, RDD estimates are no more likely to be significant than RCT estimates. The results of this first exercise are only suggestive. It may be that DID and IV methods are, for some reason, more likely to be applied in fields or to research questions where real underlying relationships are more prevalent. We discourage this view by showing the robustness of the pattern to inclusion of controls for a broad range of articles' and authors' characteristics, and journal fixed effects.

Second, we apply a caliper test. The above probit estimates suggest that IV and DID are more likely to reject the null hypothesis, but because they are estimated on the full range of p-values they have little to say about the extent of p-hacking. For instance, published studies relying on RCT or RDD may be more likely to report tightly-estimated zeros. In contrast, caliper tests focus on the pattern of p-values observed within a narrow band around arbitrary statistical significance thresholds (Gerber and Malhotra, 2008a). Caliper tests hypothesize that there should not be bunching at the significance thresholds since sampling distributions should reflect continuous probability distributions. Using this method, we find that the proportion of tests that are marginally significant is 7 to 10% higher for studies using IV than those using RCT and RDD articles. We also provide weak evidence that the

proportion of test statistics that are marginally significant is higher for DID than for RCT and RDD articles.

Third, we use the methodology developed and applied in Brodeur et al. (2016) to quantify the excess (or dearth) of p-values over various ranges by comparing the observed distribution of test statistics for each method to a counterfactual distribution that we would expect to emerge absent publication bias. We find that the extent of misallocated tests differ substantially between methods, with IV papers looking the most heavily biased. Once we have applied our methodology we assert that IV papers have approximately 55% more one-star statistics than those using RCT. These just-significant tests comprise around 20% of the unexplained test statistics for IV.

Taken as a whole, the results point to striking variations in the 'trustworthiness' of papers that use the four methods studied. Treating the bodies of published research using the methods as distinct literatures, we find the RCT and RDD corpora the most trustworthy. IV and, to a lesser extent, DID appear substantially more prone to p-hacking and selective reporting. In a further set of results, we also document a sizeable over-representation of first stage F-statistics just over the threshold of 10, suggesting p-hacking in IV is also present in the first stage. Our findings are broadly consistent with a growing literature discussing model misspecifications for IV regressions (see, for instance, Andrews et al. (2018) for a discussion on weak instruments). Using 1,359 instrumental variables regressions from 31 published studies, Young (2018) show that more than half of the statistically significant IV results depend upon only one or two outlier observations or clusters.

Our paper contributes to a discussion of the trustworthiness or dependability of empirical claims made by economics researchers (see Christensen and Miguel (2018) for a recent literature review). Using test statistics from three prestigious economics journals, Brodeur et al. (2016) provide evidence that 10 to 20 percent of marginally rejected tests are false-positives. We extend the work of Brodeur et al. (2016) in several ways by, for instance, comparing the top 5 and remaining top economic outlets. The distribution of test statistics has a two-humped camel shape for both subsamples, suggesting that pedigree does not indicate an absence of publication bias. Another important study, Vivalt (2019), investigates

the extent of p-hacking for a large set of impact evaluations. Vivalt (2019) and Brodeur et al. (2016) both provide evidence that the extent of p-hacking is smaller for RCT than for other methods. We complement these studies by including a large number of articles' and authors' characteristics in our model and by partitioning p-hacking for quasi-experimental methods; the most commonly used identification strategies in many social sciences.

Our study also contributes to a large literature documenting new ways to enhance the credibility of empirical research (Miguel et al., 2014). To some extent, our findings suggest that improved research design itself may partially constrain p-hacking, although other solutions are necessary.[1] Last, our results point to the importance of identifying and correcting publication bias (Andrews and Kasy (2017)) and that the appropriate correction is sensitive to method. They may also explain divergences documented in meta-analyses in the size and precision of estimates within a given literature (e.g., Havránek and Sokolova (2016)).

The rest of the paper is organized as follows. Section 1 details the data collection and methodology. We discuss the results in section 2. Section 3 concludes.

1 Data and Methods

1.1 Data Collection

We collect the universe of articles published by 25 top journals in economics during the 2015 calendar year. Table 1 provides the complete list of journals included in the analysis. We selected top journals as ranked using RePEc's Simple Impact Factor.[2] We excluded any journal (e.g., *the Journal of Economics Perspectives*) that did not publish at least one paper employing one of the methods under investigation.

wr In selecting our samples we followed a rule-based exclusion procedure. For each method we began by searching the entire corpus of published articles for keywords related

[1] A growing number of solutions such as data sharing and pre-analysis plans have been proposed to make research more transparent. See Blanco-Perez and Brodeur (2019) for a survey of economics journals data and code availability policies and Olken (2015) for a discussion of the pros and cons of pre-analysis plans.

[2] RePEc's Simple Impact Factor, calculated over the last 10 years. This measure uses a citation count and scales it by the number of articles in each journal. Within-journal citations are not included. Accessible at `https://ideas.repec.org/top/top.journals.simple10.html`.

to that method.[3] These keywords provide four corpora, one corresponding to each method.[4] We manually excise articles if they employed a sub-method that alters researcher freedoms. We thus remove papers that use matching (DID) and papers that use instruments as part of a fuzzy RDD, focusing exclusively on two stage least squares (IV). We also remove any papers that use a Structural Equation Model.[5] See Appendix Table A1 for an example of our data collection for a journal article. We ultimately collect statistics from 300 articles.

Journals do not contribute equally as slightly more than half of the articles come from the following seven journals: the *American Economic Journal*: *Applied Economics* and *Economic Policy*, the *American Economic Review*, the *Journal of Development Economics*, the *Journal of Financial Economics*, the *Journal of Public Economics* and the *Review of Economics and Statistics*.

From the surviving articles, we collect estimates from results tables. Our goal is to collect only coefficients of interest, or main results - we exclude any obvious regression controls or constant terms. We also exclude summary statistics, balance and robustness checks, heterogeneity of effects, placebo tests, etc. Coefficients drawn from multiple specifications of the same hypothesis are collected. All reported decimal places are collected. For DID, we collect only the interaction term, unless the non-interacted terms are described by the author(s) as coefficients of interest. For IV, we only collect the coefficient(s) of the instrumented variable(s) presented in the second stage. For papers that use more than one method, we collect estimates from each, e.g., if a paper uses both DID and IV, we collect estimates for both and add them to the relevant method's sample.[6] Although these rules make clear the majority of exclusion decisions, they are not comprehensive. In cases of ambiguity we err on the side of exclusion.

[3]For DID: "difference-in-difference*", "differences-in-difference*", "difference in difference*" and "differences in difference*". For IV: "instrumental variable*". For RCT: "randomized". For RDD: "regression discontinuity". Where * represents a wildcard in the text search, allowing for plurals to be captured with the same search string.

[4]We manually excluded articles that contained the search term but did not use one of the four methods.

[5]The AEA's RCT registry was established in 2012. Nonetheless, few RCT articles in our sample mention that they relied on a pre-analysis plan (or pre-registration). This is probably due to the time necessary to complete field experiments and the long delays from submission to publication.

[6]For field experiments with partial compliance, we add the Intention-to-Treat estimates to the RCT sample and the IV estimates to the IV sample. In our sample, only five studies used both IV and RCT methods.

To avoid inconsistencies and coding errors, test statistics were manually collected by us, the authors. Moreover, each article was independently coded by two of the three authors. This exercise allowed us to reproduce the work of one another and to make sure we only selected coefficients of interest. Note that we collected the same test statistics for the vast majority of the articles. We talked at length about the test statistics for which there was initially a disagreement. In the end, we collected the same tests or easily reach an agreement for 98.7% of test statistics.[7] We keep the remaining tests in our sample, but check that they do not drive our main findings in Section 2.

All the test statistics in our sample are two-tailed tests. The majority (91%) of test statistics are reported as coefficients and standard errors, with a minority presented directly as t-statistics (6%) or p-values (3%). Because degrees of freedom are not always reported, we treat coefficient and standard error ratios as if they follow an asymptotically standard normal distribution. When articles report t-statistics or p-values, we transform them into the equivalent z-statistics produced by the coefficient and standard error ratios. Note that our conclusions prove robust to the way authors report their results.

For each article, we also collected information about the academic affiliation of the authors, the number of authors and the number of Google Scholar and Web of Science citations (as of December 2018). In addition to articles' characteristics, we collected additional information for the authors in our sample. We managed to get the following information from curriculum vitae for about 96% of the authors in our sample: year and institution the PhD was earned, gender, the author's academic rank (e.g., assistant professor) as of 2015 and whether the author was an editor of an economic journal.

1.2 Descriptive Statistics

We ultimately collect a total of 12,213 test statistics. On average, we collect 29 estimates per article for DID and 22 estimates per IV article. RCT and RDD offer 55 and 69 coefficients per article. In our analyses we include article and table weights to prevent articles and tables

[7]Most of the tests for which there was a disagreement were differences-in-differences. We sometimes could not agree on whether the non-interacted terms were coefficients of interest.

with more tests from having a disproportionate effect. Table 1 summarizes the contribution of each method to the sample. Overall, the frequency of tests is roughly comparable across methodologies, with shares of the dataset for DID, IV, RCT and RDD at 24%, 22%, 29% and 25% respectively.

Table 2 provides descriptive statistics for articles' and authors' characteristics. The unit of observation is a test statistic. In our sample, about half of the articles have at least three authors, whereas less than 25% are solo-authored. Approximately three-quarters of the authors are males and the mean academic year of graduation is 2003–2004.[8] A rough categorization of institutions into top and non-top[9] reveals that approximately 40% of the articles had at least one author from a top institution. The figure increases to 65% for at least one author who graduated from top a institution. Last, the average number of Web of Science citations is 17 (std. dev. of 17).

A decomposition by methods reveals that authors using RDD (RCT) earned their PhD more (less) recently than authors using IV and DID. Female authors are more likely to rely on RCT, and less so on DID. Moreover, authors from top institutions are more (less) likely to use RCT (RDD) than DID and IV. We include in our model authors' and articles' characteristics to control for these compositional differences.

1.3 Distribution of Tests

Figure 1, panel A presents the raw distribution of z-statistics in our sample. This figure displays barcharts of z-statistics. Each bar has a width of 0.10 and the interval $z \in [0, 10]$ was chosen to create a total of 100 bins. Accent lines are provided at the conventional two-tailed significance levels. The distribution presents a two-humped camel shape with a first hump with low z-statistics and a second hump between 1.65 and 2.5. The distribution exhibits a local minimum around 1.35, suggesting *misallocated* z-statistics. About 57, 49 and 34% of test statistics are respectively significant at the 10, 5 and 1 percent levels. This

[8] Nine authors had not completed their PhD at the time of publication.

[9] The following institutions are coded as top: Boston U, Brown, Chicago, Columbia, Cornell, Dartmouth, Harvard, LSE, Michigan, MIT, Northwestern, NYU, Princeton, UC Berkeley, UCLA, UCSD, UPenn, Stanford and Yale.

is somewhat in line with Brodeur et al. (2016) who documented that approximately 54% of tests were statistically significant at the 5 percent level in three top economics journals.

The number of tests per article varies considerably in our sample (mean of 40 and std. of 43). To alleviate this issue, we weight each test statistic using the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article. This weighting scheme puts equal weights on each article and equal weights on each table of a same article. Panel B illustrates the weighted distribution of tests. The shape of the distribution remains similar to the unweighted distribution. See Appendix Figure A1 for a direct comparison of the raw and weighted distributions.

Figure 1, panels C and D split the full sample of published z-statistics depending on the rank of the journals. Panel C restricts the sample to the "Top 5",[10] while panel D shows the distribution of tests for the remaining journals. Both distributions present a two-humped shape, although the number of tests with high p-values is larger for top 5 than for non-top 5 publications. We formally control for journal ranking in our analysis.

Figure 2 displays barcharts of z-statistics for each of the four methods. (See Appendix Figure A2 for the weighted distributions.) We create Z-curves by imposing an Epanechnikov kernel density (also of width 0.10). A kernel smooths the distribution, softening both valleys and peaks. In Figure 3, we plot the same Z-curves into a single panel.

The shapes are striking. The distributions for IV and DID present a global and local maximum around 2 (p-value of 0.05), respectively. DID and IV seem to exhibit a mass shift away from the marginally statistically insignificant interval (just left of $z = 1.65$) into regions conventionally accepted as statistically significant, indicative of p-hacking. The extent of p-hacking seems to be the highest for IV with a sizable spike and maximum density around 1.96. The distributions for IV and DID are increasing over the interval $[1.5 - 2]$ and has the largest proportion of tests that are statistically significant at the 10 to 1 percent level.

In stark contrast, RDD presents a more or less smooth and monotonically falling curve with a maximum density near 0. The distribution for RCT is somewhat similar to RDD with most p-values lower than 0.5. The extent of p-hacking in the strands of literature using

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

these methods seems much more limited than those using IV and DID.

Visual inspection of the patterns suggests two important differences between these two groups of methods. First, many RCT and RDD studies report negative results with large p-values for their main estimates, whereas IV and DID studies typically reject the null hypothesis. Second, DID and IV are more likely to report marginally significant estimates than RCT and DID, suggesting that the extent of p-hacking is related to methods. An alternative explanation is that editors' and referees' preferences for negative results may differ by method. We probe this further in the regression analysis that follows.

We rely on the additional data we collected on articles and authors to check whether these patterns are visible across different subsamples. Appendix Figures A3-A9 display the decompositions by methods along the following characteristics: citations, number of authors, institution rank, PhD institution rank, years of experience since PhD, editor of an economic journal and gender, respectively.

This bird's-eye-view across subsamples indicates that distributions of z-statistics vary with these features, but that our main finding is apparent for all these subsamples: the distributions of tests for DID and IV present a two-humped camel shape, while the distributions of tests for RCT and RDD present a more or less smooth and monotonically falling curve. Of note, though, the spike at about $z = 2$ is particularly striking for solo-authored and non-top affiliation authored IV studies.[11]

1.4 Methodology

In this study, we are interested in documenting the number of test statistics within a narrow band for each method, but also the whole distribution of tests. On the one hand, a local analysis could shed light on the extent of p-hacking if an unusually large number of tests are just over the critical values (e.g., Gerber and Malhotra (2008b)). On the other hand, looking at the whole distribution allows us to check whether authors may be more or less likely to "file-drawer" negative results depending on methods. For instance, a researcher

[11] Another interesting observation is that there are many well-estimated zeros (and virtually no bunching at about $z = 2$) for RCTs with at least one author at a top institution.

may abandon a DID approach that fails to yield a positive result, but submit an RCT not rejecting the null. This behavior would lead to a large number of tests with high p-values.

We document the differences in selective reporting by method in three stages. First, we rely on probit regressions and compare the quasi-experimental methods from the baseline distribution obtained from RCTs. Second, we rely on the caliper test and restrict the sample to a narrow band around statistically significant thresholds. Third, we plot the distribution of z-statistics for the four methods and compare it to plausible counterfactual distributions.

For the first and second exercises we estimate the following equation:

$$Significant_{ij} = \alpha + \beta_j + X'_{ij}\delta + \gamma DID_{ij} + \lambda IV_{ij} + \phi RDD_{ij} + \varepsilon_{ij} \quad (1)$$

where $Significant_{ijf}$ is an indicator variable that estimate i is statistically significant in journal j. We include indicators for individual journals. Our results hold within-journal. We report marginal effects from probit regression throughout. Standard errors are clustered at article level.[12]

One potential issue with our identification strategy is editors' and referees' preferences for negative results may differ by method. Though plausible, the inclusion of journal fixed effects in our model leads us to believe this is not driving our results. Another potential issue is that the extent of p-hacking in a method could be related to the characteristics of the authors that use that method. We tackle this issue by including the term X_{it} in our model. This vector includes indicators for reporting methods, i.e., p-values or t-statistics, table ordering, the natural log of one plus Web of Science citations and a broad range of authors' characteristics.

A third potential issue is that different methods may answer different questions, leading to different rejection rates. We tackle this issue with the caliper test. For this exercise, we restrict the sample to a narrow band around a statistical significance threshold. While research methods can have different rejection rates, they should not lead to more or less estimates that are marginally significant.

[12]Clustering by journal yields very similar conclusions. Estimates available upon request.

For the caliper analysis, we use a baseline window of ± 0.5. This means that we restrict the sample to $z \in [1.46, 2.46]$ when the dependent variable is whether the test is statistically significant at the 5 percent level. We later explore the sensitivity of our findings to window width.

The main criticism of caliper methods is that bunching below/above statistical thresholds reflects good prior knowledge of the sample size necessary to obtain a marginally significant estimate.[13] We think it is unlikely we are vulnerable to this critique for two reasons. First, it is in RCTs that researchers are most able to choose their sample size based on their power calculations, not in IV or DID. Second, sample size for the articles in our sample is much smaller for RCTs than for the other methods, especially DID. So if the bunching reflects good priors and power calculations, then the bunching should be much more present in RCTs.

2 Results

2.1 Probit Estimates: Whole Sample

Table 3 presents estimates of Equation 1 where the dependent variable indicates whether a test statistic is statistically significant at the 5 percent level. The coefficients presented are increases in the probability of statistical significance relative to the baseline category (RCT). We report standard errors adjusted for clustering by article in parentheses. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations. (See Appendix Table A2 for the unweighted estimates.) This weighting scheme is used to prevent tables (and articles) with many test statistics to be overweighted.

In the most parsimonious specification, we find that DID and IV estimates are about 15% more likely to be statistically significant than a RCT estimate. Our IV estimate is statistically significant at the 1 percent level, while the DID estimate is statistically

[13] See Ioannidis et al. (2017) for an investigation of statistical power and bias in economics. They document that many research areas in economics have nearly 90% of their results under-powered.

significant at the 2 percent level. In contrast, RDD estimates are *not* statistically more likely than RCT estimates to be statistically significant.

In columns 2–4, we enrich our specifications with journals fixed effects and articles' characteristics. In column 2, we include an indicator for whether the estimate is in an article published in a top 5 journal. In column 3, we include individual journal fixed effects. In column 4, we control for the natural log of Web of Science citations[14] and add indicators for reporting t-statistic or p-values directly. In order to control for table ordering, we also include a variable indicating the table number in which the estimate is presented in the journal article. We report estimates for the control variables in Appendix Table A3.[15]

Our IV estimates remain statistically significant at the 1 percent level across specifications and range from 13% to 14%. Our DID estimates are significant at the 5 percent level and range from 14% to 16%. On the other hand, the RDD estimates are consistently small and statistically insignificant.

In columns 5 and 6, we add to the model the following authors' characteristics: solo-authorship, the average years of experience since PhD, average experience-squared, share of female authors, (column 5), share of authors from a top institution, share of authors who graduated from a top institution and an indicator for whether at least one of the authors was an editor of an economic journal at the time of publication (column 6). The inclusion of these additional controls has no effect on the size and significance of the DID, IV and RDD estimates. This result suggests that it is unlikely that compositional differences by methods are driving our findings.

Appendix Tables A4 and A5 replicate Table 3 for the other conventional significance levels. For the 1.65 cutoff, the estimates are statistically significant throughout for IV (DID) in comparison to RCT and range from 13% to 15% (12% to 13%). In contrast, we find no evidence that methods are related to the 2.58 two-sided cutoff.

Overall, these findings provide evidence that within top journals in economics, the likelihood of a null hypothesis being rejected is related to the underlying research method. In

[14] Using Google Scholar instead of Web of Science to track citation counts yields very similar results.

[15] We do not find that estimates published in top 5 journals are more or less likely to be statistically significant than estimates outside of the top 5, conditional (or not) on method used. On the other hand, citations are positively associated with the likelihood to reject the null hypothesis.

the next subsection, we test whether these patterns hold when we restrict the sample to marginally significant and insignificant tests.

2.2 Caliper Test

We now rely on the caliper test. This test compares the number of estimates in a narrow range below and above a statistical significance threshold. If there is no manipulation we would expect the number of test statistics that fall just above an arbitrary threshold, such as 1.96, to be very similar to the number that fall just below.

Table 4 reports the estimates for the 5 percent significance level. The table structure is the same as Table 3. The only difference is that we restrict the sample to $z \in [1.46, 2.46]$. Our sample size decreases to 2,922 observations. We report estimates for the control variables in Appendix Table A6.

Our probit estimates suggest that IV articles are significantly more likely to report marginally significant test statistics at the 5 percent level than RCT and RDD articles. Results are statistically significant in all columns and range from 10% to 15% (RCT is the baseline). Similarly, we find that DID articles are more likely to report marginally significant tests. Our estimates are significant at conventional levels in all columns, suggesting that DID studies are about 15% more likely to report marginally significant tests than RCTs.

As before, we present estimates for the other conventional levels in Appendix Tables A7 and A8. We also apply a window of ± 0.5 for the other two significance levels. We confirm that IV articles are significantly more likely to report marginally significant tests at the 10 percent level than RCTs. The estimates are all significant and range from 10% to 15%. Interestingly, RDD estimates are significantly *less* likely to report marginally significant tests at the 10 percent level than DID, IV and RCT articles. There is no significant differences between DID and RCT papers. Last, we do not find that the extent of p-hacking varies by methods for the 1 percent significance threshold. Once this high level of significance has been reached, differences between methods become small and statistically insignificant.

Appendix Tables A9 and A10 show that our caliper findings for the 1.96 and 1.65 cutoffs are robust to other windows. More precisely, columns 1–5 restrict the sample to tests within

the following windows: 1.96 (or 1.65) ±0.40, ±0.45, ±0.50, ±0.55 and ±0.60, respectively. The point estimates for DID and IV are all positive and significant at conventional levels in most columns. As expected, the smaller the window, the more imprecise the estimates are.

In summary, our results provide suggestive evidence that the extent of p-hacking differs by methods and levels of significance. Researchers using IV (and DID to some extent) might be tempted to choose a slightly more "significant" specification to just pass the significance thresholds at the 10 or 5 percent levels. On the other hand, researchers relying on RCT and RDD may be either less inclined to p-hack and/or have less discretion to inflate their z-statistics.

2.2.1 Robustness Checks

As a validation to our probit analysis, we compare our estimates with probit regressions to identical specifications using logistic models. We show that our main findings are robust to the use of a logit specification in Appendix Tables A11 and A12. The structure of the table is the same as in Tables 3 and 4. Our point estimates are virtually the same and remain statistically significant at conventional levels.

Appendix Tables A13 and A14 report additional robustness checks for the whole sample and the caliper test exercise, respectively. We first check whether our findings are robust to coding/data collection methods. As mentioned before, we replicated the work of each other and ended up collecting the same tests for the vast majority of research articles. In columns 1–3, we drop the articles for which we could not easily reach an agreement on which tests to select. The dependent variables are dummies for whether the test statistic is significant at the 1, 5 and 10 percent level, respectively. The point estimates are very similar and suggest that IV articles are significantly more likely to report marginally significant test statistics at the 5 and 10 percent level than RCT and RDD articles.

In columns 4–6, we test whether the main findings are not driven by journal articles relying on multiple methods. About 14 percent of the tests in our sample are in an article using multiple methods. This includes, for instance, journal articles using both DID and RDD to answer a research question, and a combination of RCT and IV for papers with

partial compliance. Excluding these papers has no effect on our main conclusions. For both the caliper exercise and the whole sample, we find that DID and IV articles report more marginally significant tests at the 5 percent level than RCT and RDD articles. The estimates are statistically significant at the 5 percent level.

Thus far, we have relied on all journals in our sample. As a robustness check, we explore the sensitivity of our results to the omission of a subset of journals. In Appendix Table A15, we check whether omitting a set of journals within an economic field in the analysis affects the main results. We create dummies for the following "fields" in our sample: top 5 , general interest (not top 5), development, experimental, finance, labor, macroeconomics, public and urban. Hence, we tabulate the estimates of nine probit regressions. As with our prior estimates, this sensitivity test suggests that tests relying on RCT and RDD papers are less likely to reject the null hypothesis than IV and DID tests. All our estimates are statistically significant and range from xx to xx for IV and from xx to xx for DID (RCT is the omitted category). This suggests that our findings are unlikely to be driven by specific fields or journals.

2.3 Excess Test Statistics

In this subsection, we detail our third strategy that compares the differences of an observed test-statistic distribution to a monotonically decreasing one. It is loosely based on a ratio methodology introduced in Brodeur et al. (2016), however the method we apply makes fewer assumptions and requires less structure. The cost of fewer assumptions is that our main results are necessarily relative to a baseline.

The methodology follows a simple recipe and needs few ingredients. First, we require a proxy for the unknowable distribution of possible test statistics absent researchers' or reviewers' influences. Second, the observed distribution of published test statistics. From these two distributions we generate the difference of observed to underlying or expected test statistics. The area below these difference distributions, which corresponds to the area between observed distribution and the expected, is the mass of unexplained test statistics. Third, these masses are compared to those generated by the RCT difference distribution.

2.3.1 Methodology

We first look for a reasonable input distribution for all possible test statistics. Consider a working paper in the universe of all those possible. This working paper tests a unique hypothesis, and we denote by z the absolute value of its test statistic. Over all working papers, the distribution of these z form the basis of an input distribution - what a test statistic distribution would present as absent any researcher or publication influences. Since the natural distribution of tests prior to publication bias or p-hacking is unobserved, we make the same modeling decision as in (Brodeur et al., 2016) and use the Student-t distribution with one degree of freedom and the Cauchy(0,0.5) distribution.[16] See Appendix Figures A10 and A11. The primary feature for any reasonable input distribution is a probability density function that is weakly monotonically decreasing from zero.[17] Our results, particularly when comparing unexplained masses between methods, are not sensitive to the choice of input distribution or weighting scheme.

The second ingredient is the observed test statistic distribution. To ensure the extreme tails do not drive our results in the locality of significance thresholds, we normalize the area below each observed distribution (and the input distributions) to one in the interval $z = [0, 10]$. For each of the four methods, we generate the difference of the observed distribution to the input distribution. A negative number implies that there are 'missing' test statistics for that value of z whereas a positive number implies an excess of test statistics. See Appendix Figures A12-A15. We then sum these missing test statistics (which is the area below/above a difference curve) by significance region. We partition the interval $z = [0, 10]$ using the common significance thresholds for a two-tailed test at the 10%, 5% and 1% levels. Colloquially we can refer to these as the insignificance, one-star, two-star, and three-star regions.

[16] (Brodeur et al., 2016) justify this decision by demonstrating how well the observed and input distributions reflect each other in the extreme significance region of $z = [10, 20]$, where distortion is less likely to occur.

[17] Tests of correlations between variables chosen from a pool of uncorrelated processes would follow this distribution. When looking to our observed distributions (regardless of method) we see tails markedly larger than could be generated by the Gaussian family of distributions. We also use the Cauchy distribution (with location parameter zero and scale parameter 0.5), a relatively fat-tail ratio distribution. This seems natural in our case because the ratio of two normal distributions follows the Cauchy distribution.

Third, we compare the unexplained mass of test statistics for each method to a common baseline. To apply our methodology, we assume that articles using one of the quasi-experimental methods are treated equally as an RCT in the different significance regions - that an article using an IV and achieving the same significance level is just as likely to be published in our wide set of journals as an article using an RCT. Wherever this assumption is tenuous we caution interpretation of our results. For example, many RCTs use power calculations creating a possible dearth in test statistics above three stars. This would be reflected in very large unexplained masses above $z = 2.58$ for DID, IV and RDD. In the same token, if we believe that RCTs are differentially published with the lowest levels of significance, such as in the one star region, then we would expect there to be too few test statistics in the interval $z = [0, 1.65)$ and too many in the intervals $z = [1.65, 10]$. We see both of these artifacts in our results and thus focus most of our discussion on results in the two-star interval.

2.3.2 Misallocation of Tests

In Table 5, we present the results of our differences strategy. The table is made of four panels, each using a different input distribution or weighting strategy. The top panel uses the Student-t distribution with a single degree of freedom and does not apply weights to the distribution of observed test statistics. In panel B, we present the same exercise with the same input while applying article $\times$ table weights to the observed test statistics. This reflects the possibility that a reader may consider all tables in all articles evenly when consuming research. Panels C and D reflect the top two panels while using the Cauchy distribution with a location parameter of zero and a scale parameter of 0.5.

The first result in the top panel suggests that 23% of DID and 29% of IV test statistics are 'missing' from the interval $z = [0, 1.28)$. This estimate of test statistics corresponds directly to the area between the DID-input difference curve and the zero line in the interval $z = [0, 1.28)$ in Appendix Figure A12. We then show the extent of misallocation for the intervals $z = [1.28, 1.65)$, $z = [1.65, 1.96)$, $z = [1.96, 2.58)$ and $z = [2.58, 3.29)$. We find a surplus of 2.3% of DID and 3% of IV test statistics in the one star significance interval,

$z = [1.65, 1.96)$. In the two star significance interval, 8.3% and 9.8% of DID and IV of test statistics are misallocated, respectively. The remaining test statistics can be found above the common three star threshold and below just below the one star threshold. Missallocation in the one and two star significance intervals is relatively smaller for RCT and RDD studies. For instance, we that 5.4% of RCT and 5.9% of RDD test statistics are found in the two star significance interval.

Overall, 10.6, 12.8, 8.2 and 7.6% of tests are misallocated in the one and two star intervals for DID, IV, RCT and RDD, respectively. One major note here is that despite making up very little of the test statistic interval, these just significant regions contain a large amount of excess test statistics. In contrast, the extent of misallocation is very small for all four methods for the interval $z = [1.28, 1.65)$, suggesting much less 'inflation' of test statistics for insignificant estimates.

In columns 5–7, we use the RCT test statistics distribution as a baseline for our analysis. In doing so, we no longer assume we have correctly identified the underlying test distribution (only that RCTs better approximate whatever that may be) and we gain guidance on how much excess mass there should be. We present the mass ratio of each quasi-experimental method to RCT test statistics by significance region. In the non-significant region, DID has over double the missing test statistics as RCT. In the adjacent column, we show that IV is missing 2.5 times as many. Next to that, RDD is only missing 55% more test statistics than the RCT baseline.

In the just significant regions, we find the main results of this section. DID has 83% of the expected mass of test statistics in the one star region, and 54% more in the two star region. IV fares worse, with 8% more test statistics with one star and 83% more statistics with two stars than expected. RDD articles fare quite well - presenting 39% less test statistics in the one star region and 9% more tests in the two star region relative to RCT. All methods publish significantly more three star results than RCTs. This is likely due to power calculations on the part of RCTs rather than a fault common to the three other methods.

In the second panel, we repeat the exercise using the Student-t distribution of degree one

but use article × table weights to adjust the observed test statistic distributions. Although when taken to the extreme this implies that a single table article is as equally weighted as an article with many tables and hypotheses (and that those hypotheses are all equally weighted within that article), we show that even this conservative weighting scheme does not change our results.

In the third and fourth panels, we repeat the same exercise, using the Cauchy(0,0.5) distribution. There are small differences between using the Student-t and Cauchy distributions. Comparing the top panel to the third (both unweighted) we see that the one and two star misallocated masses increase for all methods. Overall, 14.5, 16.8, 12 and 11.4% of tests are now misallocated in the one and two star intervals for DID, IV, RCT and RDD, respectively. Adding article × table weights (moving to the fourth panel) has no effect on our main conclusions.

Depending on the input distribution and weighting scheme, we estimate that between 12.8–16.8% of marginally significant tests are misallocated for IV in comparison to 7.6–11.4% for RDD. This translates into about 75% more statistics with two stars for IV than for RDD (and RCT).

2.4 Instrumental Variables: F-Statistics

Patterns in the data discussed in the previous sections are consistent with p-hacking for IV studies. Our attention so far was limited to the second stage estimate. We now study the first stage, and more specifically document the distribution of F-statistics for IV articles in our sample. The first stage F-statistic is typically used in IV papers to test the hypothesis that the instrument(s) is unrelated the endogenous regressor, i.e., to test the strength of the instrument.

Using F-statistics from 17 papers published in the *American Economic Review*, Andrews et al. (2018) document the extent of weak instruments in published IV studies. Andrews et al. (2018) find that weak instruments are frequently encountered and that virtually all published papers reported at least one first-stage F-statistic. While our sample somewhat overlaps, it is larger and covers more journals. Interestingly, F-statistics were reported in

only two-thirds of IV papers in our sample. On average, there were nine F-statistics (std. dev. of 14) per paper. Some authors solely reported one F-statistic in a footnote or the Online Appendix, whereas others reported F-statistics for every IV regression. For this reason, we weight each article equally for our analysis.

Figure 4 shows the distribution of F-statistics reported in specifications over the interval $[0, 25]$. (See Appendix Figure A16 for $F \in [0, 75]$.) We truncate above at 25 for visibility. Our sample includes 717 F-statistics, of which 373 (566) are smaller than 25 (75). We first confirm Andrews et al. (2018)'s finding that some IV studies are in a range in which weak instruments are generally a concern. But we are more interested in whether there is bunching at 10. Most studies mentioned Stock and Watson's recommendation (or more generally the problem of weak instruments) that first-stage F-statistic(s) should be larger than 10. This suggests that authors effectively used this threshold.

We find that the distribution has a maximum density near 10. There is a sizeable under-representation of weak instruments relatively to F-statistics just over the threshold of 10 but also to (very) large F-statistics. Overall, these results indicate that both the first and second stages for IV studies present a misallocation of statistics/tests, with an over-representation of marginally significant statistics/tests.

3 Discussion

Aided by access to better data and advancements in theoretical econometrics, design-based research methods are credited as the primary catalyst for a 'credibility revolution' in economics (Angrist and Pischke, 2010). Our analysis suggests that not all design-based research methods are created equally when it comes to the credibility of empirical results. Our findings provide evidence that published studies using DID and (in particular) IV are more p-hacked than RCT and RDD. We believe this to be roughly consistent with an unsaid hierarchy in the profession, which often regards RCT as a gold standard and IV with skepticism.

One limitation of our study is that we cannot rule out that research methods may be better suited to answer different research questions which may have different rejection rates.

We have addressed this using article meta-data, however a more rigorous methodology warrants future research. While these differences may partly explain the large (small) number of RCT and RDD (IV and DID) studies with high (low) p-values, we argue it cannot explain the extent of p-hacking by method. In other words, different research questions may lead to different rejection rates, but this should not be only for estimates that are marginally significant.

In terms of recommendations, our findings suggest that improving research design may enhance the credibility of economic research. Although our study offers no means by which to improve research design, the evidence we present here does suggest that RDD is less prone to p-hacking than IV. This is encouraging as younger researchers in our sample were more likely to rely on RDD.[18]

[18]In our sample, the mean academic year of graduation is 2003–2004. Authors of IV articles graduated about three years prior to those using RDD.

References

Abadie, A. (2018). Statistical Non-Significance in Empirical Economics. National Bureau of Economic Research, Working Paper 24403.

Andrews, I. and Kasy, M. (2017). Identification of and Correction for Publication Bias. National Bureau of Economic Research, Working Paper 23298.

Andrews, I., Stock, J., and Sun, L. (2018). Weak Instruments in IV Regression: Theory and Practice. Mimeo: Harvard University.

Angrist, J. D. and Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics. *Journal of Economic Perspectives*, 24(2):3–30.

Blanco-Perez, C. and Brodeur, A. (2017). Publication Bias and Editorial Statement on Negative Findings. Pre-Analysis Plan: https://osf.io/mjbj2/.

Blanco-Perez, C. and Brodeur, A. (2019). Transparency in Empirical Economic Research. mimeo: University of Ottawa.

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.

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

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

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.

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. (2017). Unbiased Publication Bias. 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. 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.

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. (2016). Do Consumers Really Follow a Rule of Thumb? Three Thousand Estimates from 130 Studies Say "Probably Not". Czech National Bank, Research Department, number 2016/08.

Henry, E. (2009). Strategic Disclosure of Research Results: The Cost of Proving your Honesty. *Economic Journal*, 119(539):1036–1064.

Ioannidis, J. P. (2005). Why Most Published Research Findings Are False. *PLoS medecine*, 2(8):e124.

Ioannidis, J. P., Stanley, T. D., and Doucouliagos, H. (2017). The Power of Bias in Economics Research. *Economic Journal*, 127(605):F236–F265.

Leamer, E. E. (1983). Let's Take the Con Out of Econometrics. *American Economic Review*, 73(1):pp. 31–43.

Leamer, E. E. and Leonard, H. (1983). Reporting the Fragility of Regression Estimates. *Review of Economics and Statistics*, 65(2):pp. 306–317.

McCloskey, D. N. (1985). The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests. *American Economic Review: Papers and Proceedings*, 75(2):201–205.

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.

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

Panhans, M. T. and Singleton, J. D. (2017). The Empirical Economist's Toolkit: from Models to Methods. *History of Political Economy*, 49(Supplement):127–157.

Ridley, J., Kolm, N., Freckelton, R. P., and Gage, M. J. G. (2007). An Unexpected Influence of Widely Used Significance Thresholds on the Distribution of Reported P-Values. *Journal of Evolutionary Biology*, 20(3):1082–1089.

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

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22:1359–1366.

Stanley, T. D. (2005). Beyond Publication Bias. *Journal of Economic Surveys*, 19(3):309–345.

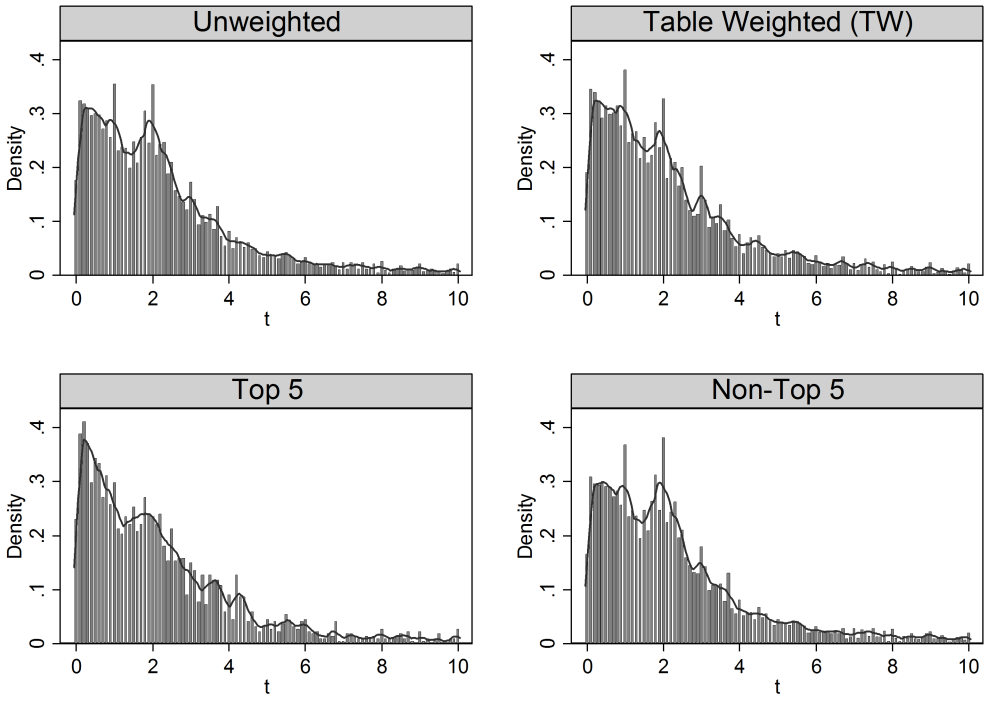
Stanley, T. D. (2008). Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection. *Oxford Bulletin of Economics and Statistics*, 70(1):103–127.

Vivalt, E. (2019). The Trajectory of Specification Searching and Publication Bias Across Methods and Disciplines. *Oxford Bulletin of Economics and Statistics*. Published Online.

Young, A. (2018). Consistency Without Inference: Instrumental Variables in Practical Application. mimeo: London School of Economics and Political Science.

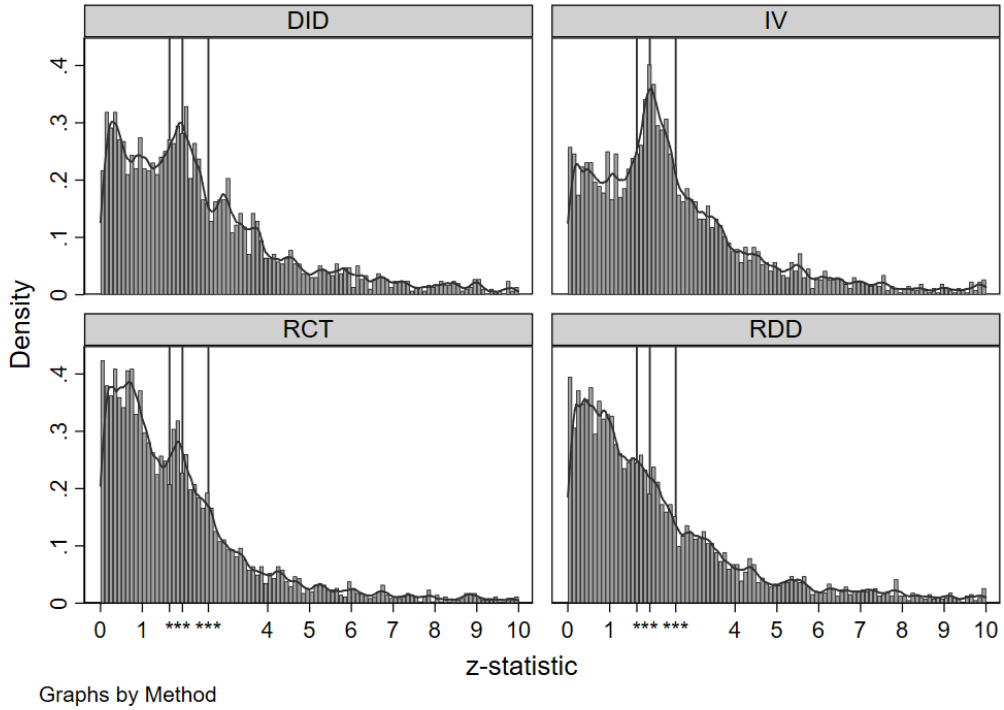
4 Figures

Figure 1: z-Statistics



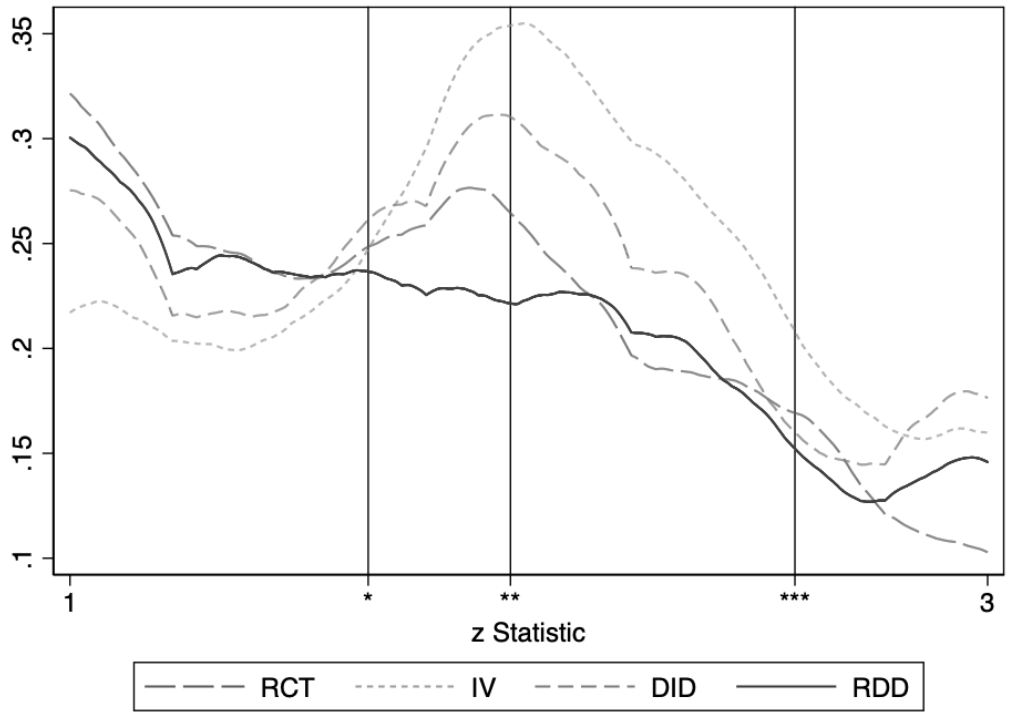
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Panel B weights each test statistic using the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article. Panels C and D restrict the sample to top 5 journals and non-top 5 journals, respectively. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure 2: z-Statistics by Method



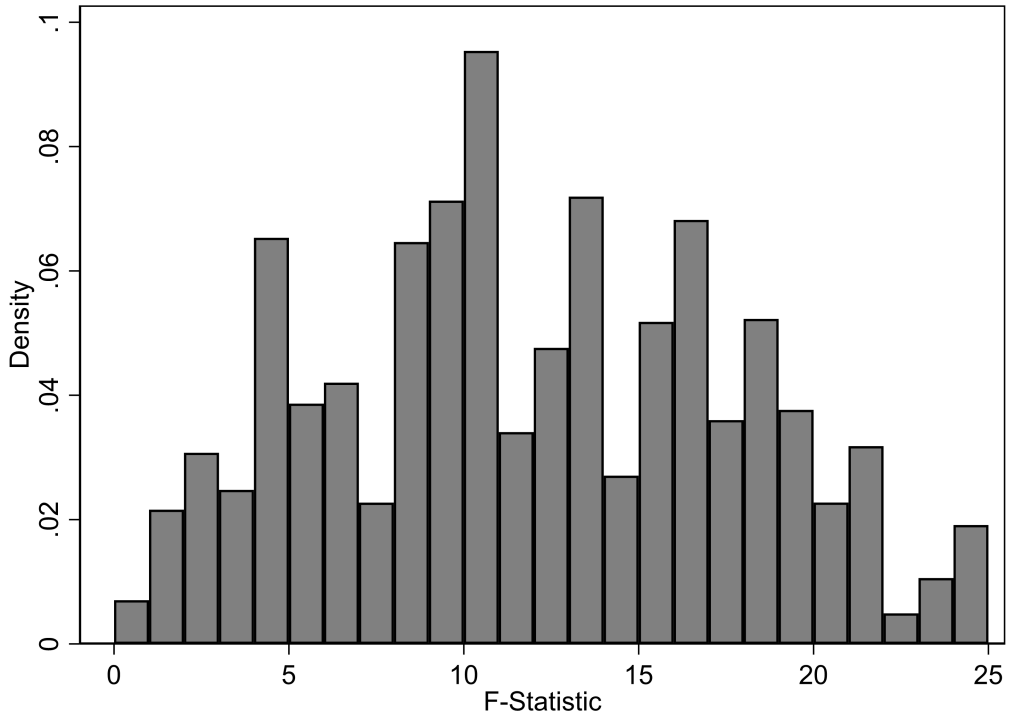
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Test statistics are partitioned by identification method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT) and regression discontinuity design (RDD). Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure 3: z-Curves by Method



Notes: This figure displays the smoothed densities (Epanechnikov) from Figure 2 for $z \in [1, 3]$. A density is displayed for each of four methods: difference-in-differences, instrumental variable, randomized control trial and regression discontinuity design. Reference lines are displayed at the conventional two-tailed significance levels.

Figure 4: Instrumental Variable: First Stage F-Statistics



Notes: This figure displays an histogram of first stage F-Statistics of instrumental variables for $F \in [0, 25]$.

5 Tables

Table 1: Summary Statistics

Journal	DID	IV	RCT	RDD	Articles	Tests
AEJ: Applied	5	8	10	2	21	1,160
AEJ: Economic Policy	11	5	2	5	21	950
AEJ: Macroeconomics		2			2	20
American Economic Review	8	11	9	2	27	1,049
Econometrica	1	2		1	4	184
Economic Journal	6	9		1	15	480
Economic Policy		1			1	6
Experimental Economics		2	1		3	29
Journal of Applied Econometrics				1	1	102
Journal of Development Economics	7	5	8	3	22	890
Journal of Economic Growth		3			3	23
Journal of Finance	3	10	3	2	16	829
Journal of Financial Economics	6	8		1	14	318
Journal of Financial Intermediation	5	5		1	11	281
Journal of Human Resources		6	2	2	10	511
Journal of International Economics	2	5			6	241
Journal of Labor Economics	3	3	5	1	11	429
Journal of Political Economy		2	2	1	5	451
Journal of Public Economics	12	6	6	7	29	1,243
Journal of Urban Economics	7	5		1	11	429
Journal of the European Economic Association	4	2	3	1	9	292
Quarterly Journal of Economics	1	5	3	3	11	474
Review of Economic Studies		3	1		4	199
Review of Economics and Statistics	8	10	8	7	29	1,130
Review of Financial Studies	11	4		3	14	493
Total Articles	100	122	63	45	300	
Total Tests	2,908	2,703	3,503	3,099		12,213

Notes: This table presents the "Top 25" journals our sample of test statistics were taken from (listed alphabetically). We identify top journals using RePEc's Simple Impact Factor: `"https://ideas.repec.org/top/top.journals.simple10.html"`. A small number of top journals did not have any eligible articles in 2015: *Journal of Economic Literature*, *Journal of Economic Perspectives*, *Journal of Monetary Economics*, *Review of Economic Dynamics*, *Annals of Economics and Finance* and the *Annual Review of Economics*. We also excluded *Brookings Papers on Economic Activity* from the sample.

Table 2: Summary Statistics: Articles and Authors' Characteristics

	Mean	Std. Dev.	Min	Max	Observations
Articles' Characteristics					
$\ln(1 + Citation)$	2.449	0.976	0	4.615	12,213
z-Statistic	0.918	0.274	0	1	12,213
t-Statistic	0.053	0.223	0	1	12,213
p-Value	0.029	0.168	0	1	12,213
Top 5	0.193	0.395	0	1	12,213
Authors' Characteristics					
Avg. Experience	10.28	6.06	0.5	38	12,185
Avg. Exp. Squared (/100)	1.424	1.793	0.003	14.44	12,185
Share Editor	0.655	0.475	0	1	12,185
Share Female	0.285	0.333	0	1	12,185
Share Top Institution	0.275	.0368	0	1	12,185
Share Top PhD Institution	0.489	0.399	0	1	12,185
Solo-Authored	0.249	0.432	0	1	12,213

Notes: This table reports articles' and authors' characteristics. Each observation is a test. For some research articles, multiple methods were used. This explains why the sum of articles for the four methods is greater than 300. The number of citations is from the Web of Science as of December 2018. Top 5 journals in economics are the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics* and *Review of Economic Studies*. The variable single-authored indicates the proportion of tests that are in solo-authored articles. The other authors' variables indicate the average years of experience since PhD, average experience-squared, the share of author(s) per article who are editors of an economic journal at the time of publication, females, affiliated to a top institution and graduated from a top institution.

Table 3: Significant at the 5% Level

	$Z > 1.96$ (1)	$Z > 1.96$ (2)	$Z > 1.96$ (3)	$Z > 1.96$ (4)	$Z > 1.96$ (5)	$Z > 1.96$ (6)
DID	0.151	0.157	0.150	0.142	0.142	0.143
	(0.061)	(0.062)	(0.066)	(0.066)	(0.062)	(0.061)
IV	0.139	0.142	0.138	0.133	0.119	0.122
	(0.050)	(0.049)	(0.053)	(0.049)	(0.049)	(0.049)
RDD	0.056	0.055	0.038	0.034	-0.001	-0.009
	(0.063)	(0.063)	(0.056)	(0.054)	(0.053)	(0.053)
Share PhD Top Institution						Y
Share Top Institution						Y
Share Female Author					Y	Y
Solo-Authored					Y	Y
Experience					Y	Y
Experience Squared					Y	Y
Editor					Y	Y
ln(Citations)				Y	Y	Y
Reporting Method				Y	Y	Y
Table Ordering				Y	Y	Y
Journal FE			Y	Y	Y	Y
Top 5		Y	Y	Y	Y	Y
Observations	12,185	12,185	12,185	12,185	12,185	12,185
Pseudo R^2	0.011	0.011	0.032	0.039	0.046	0.047

Notes: This table reports marginal effects from probit regressions (Equation (1)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

Table 4: Caliper Test, Significant at the 5% Level

	(1) $Z > 1.96$	(2) $Z > 1.96$	(3) $Z > 1.96$	(4) $Z > 1.96$	(5) $Z > 1.96$	(6) $Z > 1.96$
DID	0.159	0.155	0.158	0.155	0.148	0.149
	(0.048)	(0.048)	(0.055)	(0.055)	(0.050)	(0.050)
IV	0.106	0.105	0.124	0.152	0.133	0.131
	(0.046)	(0.046)	(0.050)	(0.051)	(0.050)	(0.050)
RDD	0.003	0.004	0.026	0.051	0.010	0.012
	(0.038)	(0.036)	(0.049)	(0.049)	(0.052)	(0.053)
Share PhD Top Institution						Y
Share Top Institution						Y
Share Female Author					Y	Y
Solo-Authored					Y	Y
Experience					Y	Y
Experience Squared					Y	Y
Editor					Y	Y
ln(Citations)				Y	Y	Y
Reporting Method				Y	Y	Y
Table Ordering				Y	Y	Y
Journal FE			Y	Y	Y	Y
Top 5		Y	Y	Y	Y	Y
Observations	2,922	2,922	2,922	2,922	2,922	2,922
Pseudo R^2	0.012	0.012	0.030	0.039	0.042	0.042
Window Width			$Z > 1.46$ & $Z < 2.46$			

Notes: This table reports marginal effects from probit regressions (Equation (1)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. We restrict the sample to $z \in [1.46, 2.46]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

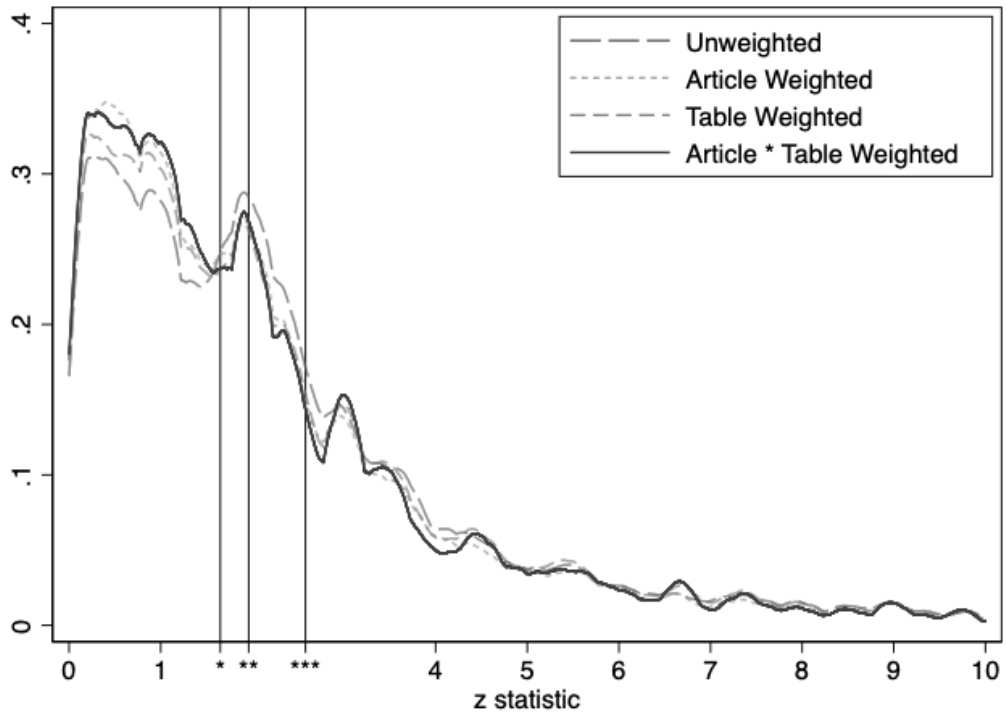
Table 5: Excess Coefficients by Significance Region

Panel A: Student(1)	DID	IV	RCT	RDD	DID/RCT	IV/RCT	RDD/RCT
Area [0,1.28)	-0.228	-0.291	-0.116	-0.180	1.968	2.504	1.549
Area [1.28,1.65)	0.001	0.001	0.010	0.007	0.097	0.069	0.718
Area [1.65-1.96)	0.023	0.030	0.028	0.017	0.826	1.075	0.610
Area [1.96-2.58)	0.083	0.098	0.054	0.059	1.543	1.825	1.092
Area [2.58-3.29]	0.050	0.054	0.027	0.040	1.871	2.001	1.483
Area (3.29,10]	0.111	0.128	0.037	0.100	2.977	3.441	2.695
Panel B: Student(1) Weighted	DID	IV	RCT	RDD	DID/RCT	IV/RCT	RDD/RCT
Area [0,1.28)	-0.213	-0.206	-0.099	-0.176	2.147	2.079	1.777
Area [1.28,1.65)	0.005	0.022	0.013	0.005	0.379	1.719	0.424
Area [1.65-1.96)	0.013	0.023	0.031	0.015	0.428	0.736	0.480
Area [1.96-2.58)	0.083	0.091	0.054	0.049	1.539	1.682	0.915
Area [2.58-3.29]	0.054	0.022	0.026	0.045	2.049	0.840	1.683
Area (3.29,10]	0.119	0.072	0.026	0.097	4.590	2.792	3.756
Panel C: Cauchy(0,0.5)	DID	IV	RCT	RDD	DID/RCT	IV/RCT	RDD/RCT
Area [0,1.28)	-0.354	-0.416	-0.242	-0.305	1.466	1.721	1.263
Area [1.28,1.65)	0.025	0.025	0.034	0.031	0.746	0.740	0.921
Area [1.65-1.96)	0.037	0.044	0.041	0.030	0.882	1.052	0.736
Area [1.96-2.58)	0.108	0.124	0.079	0.084	1.369	1.562	1.062
Area [2.58-3.29]	0.071	0.074	0.047	0.060	1.493	1.567	1.274
Area (3.29,10]	0.176	0.193	0.102	0.165	1.721	1.882	1.615
Panel D: Cauchy(0,0.5) Weighted	DID	IV	RCT	RDD	DID/RCT	IV/RCT	RDD/RCT
Area [0,1.28)	-0.339	-0.332	-0.225	-0.302	1.507	1.475	1.343
Area [1.28,1.65)	0.029	0.047	0.037	0.030	0.786	1.249	0.802
Area [1.65-1.96)	0.027	0.036	0.044	0.028	0.602	0.817	0.638
Area [1.96-2.58)	0.108	0.116	0.079	0.075	1.367	1.465	0.942
Area [2.58-3.29]	0.075	0.043	0.047	0.065	1.589	0.911	1.384
Area (3.29,10]	0.185	0.137	0.091	0.163	2.025	1.504	1.784

Notes: This table reports the difference of the observed distribution to the input distribution. A negative number implies that there are 'missing' test statistics for that value of z whereas a positive number implies an excess of test statistics. Panel A uses the Student-t distribution with a single degree of freedom. In panel B, we present the same exercise with the same input while applying article $\times$ table weights to the observed test statistics. Panels C and D reflect the top two panels while using the Cauchy distribution with a location parameter of zero and a scale parameter of 0.5. Columns 5–7 present the mass ratio of each quasi-experimental method to RCT test statistics by significance region.

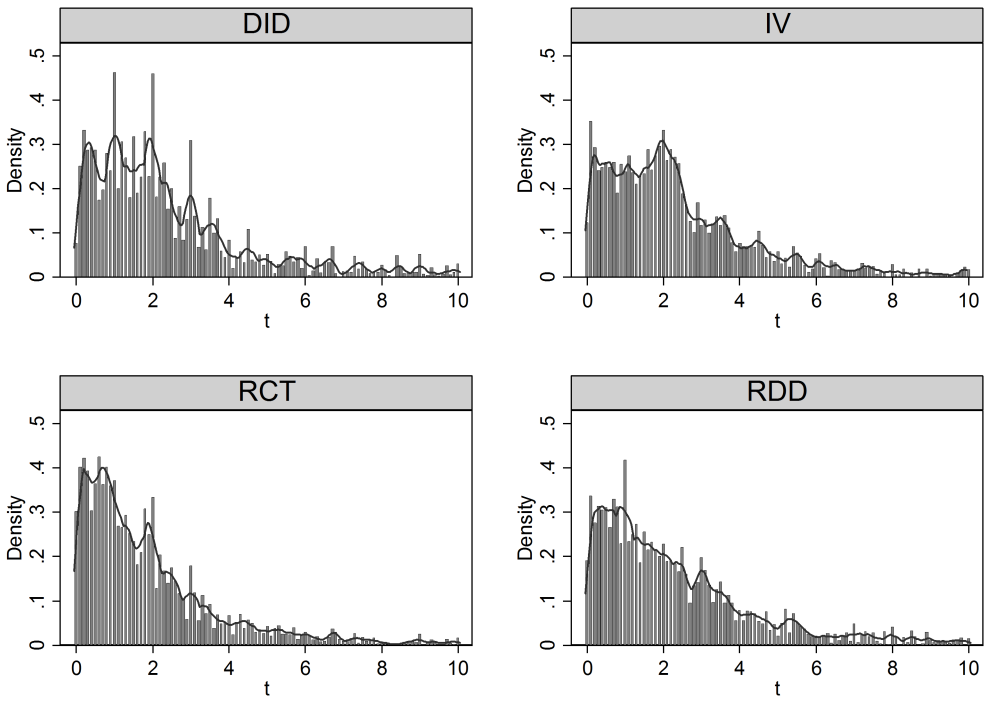
Appendices: NOT FOR PUBLICATION

Figure A1: z-Statistics by Weighting Scheme



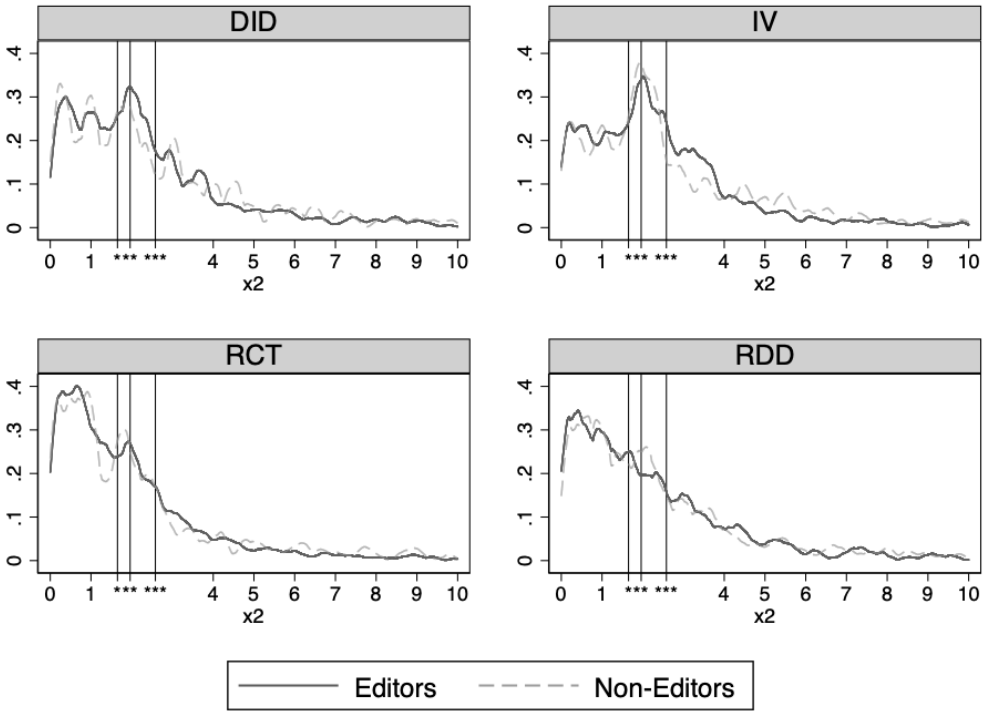
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. We present the unweighted distribution of tests, but also the weighted by article, the weighted by table and the weighted by article and table distributions. For the article and table weights, we weight each test statistic using the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure A2: z-Statistics by Method: Weighted



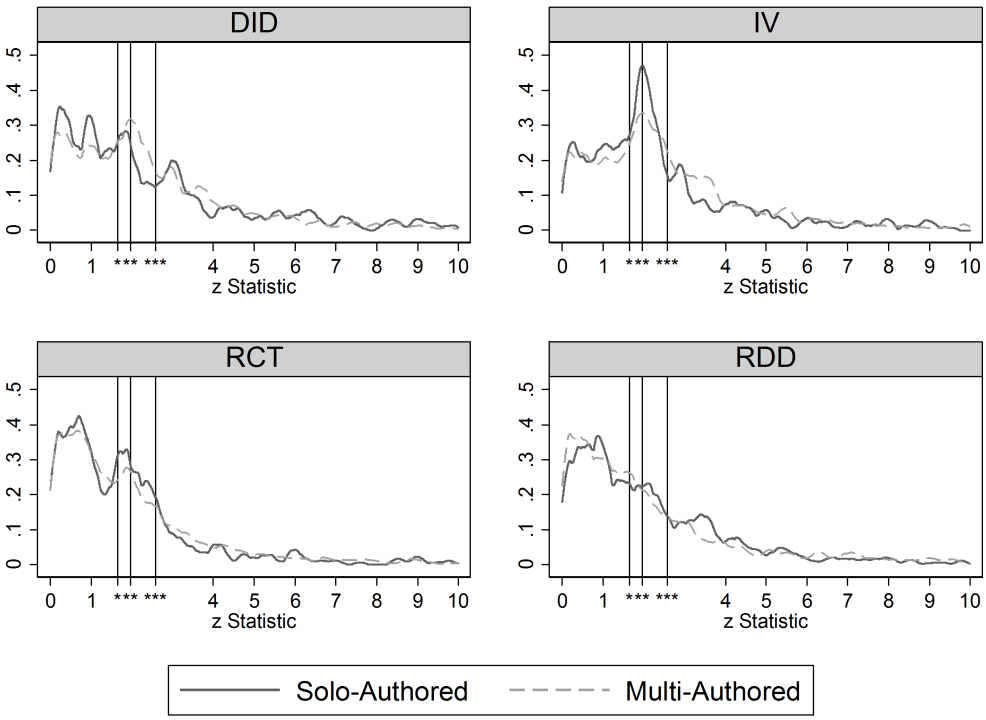
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. We weight each test statistic using the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article. Test statistics are partitioned by identification method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT) and regression discontinuity design (RDD). Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure A3: z-Statistics by Method and Citations



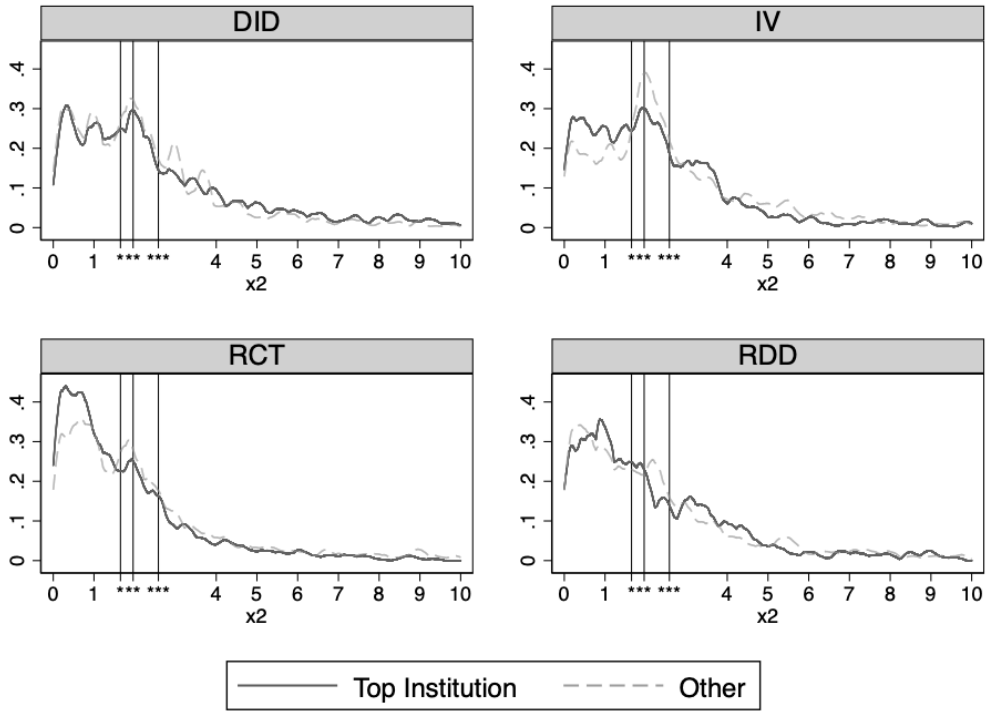
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Test statistics are partitioned by identification method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT) and regression discontinuity design (RDD). Lines in dark gray are for articles with more than the median number of Web of Science citations in our sample. Lines in light gray (dashes) are for articles with less than the median number of Web of Science citations in our sample. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure A4: z-Statistics by Method and Number of Authors



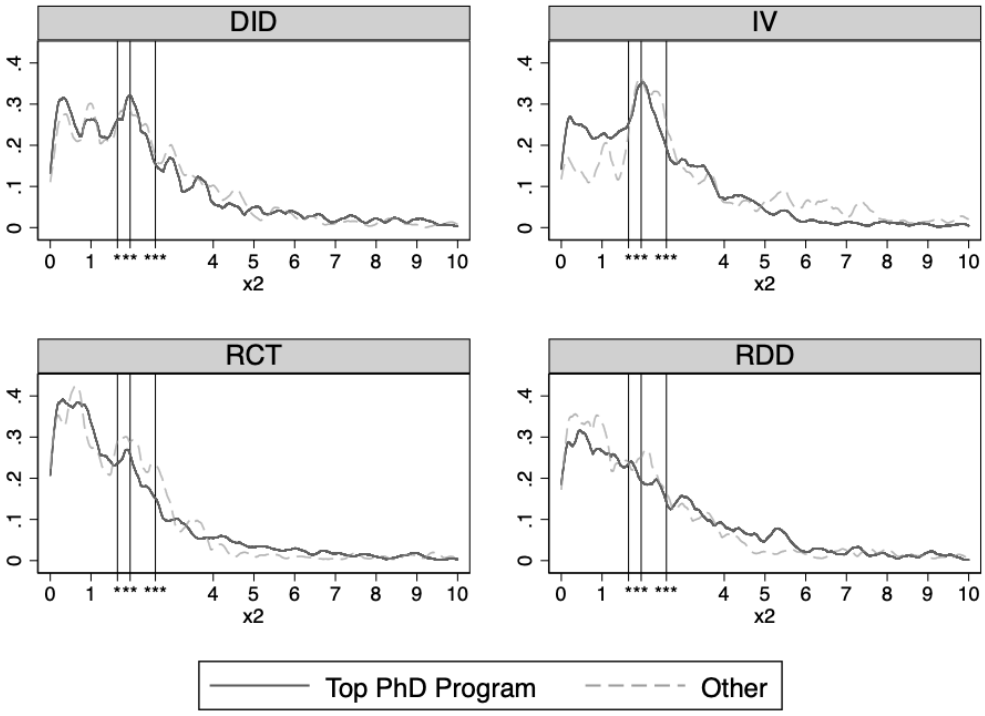
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Test statistics are partitioned by identification method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT) and regression discontinuity design (RDD). Lines in dark gray are for solo-articles. Lines in light gray (dashes) are for multi-authored articles. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure A5: z-Statistics by Method and Affiliation



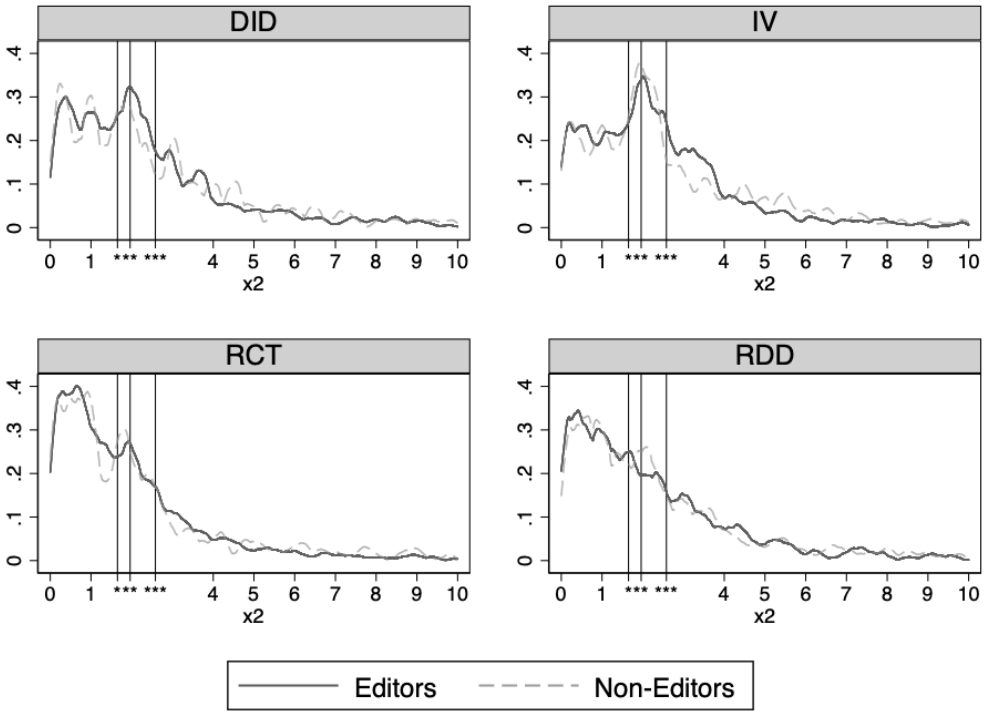
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Test statistics are partitioned by identification method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT) and regression discontinuity design (RDD). Lines in dark gray are for articles with at least one author affiliated to a top institution. Lines in light gray (dashes) are for articles with no author affiliated to a top institution. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure A6: z-Statistics by Method and PhD Institution



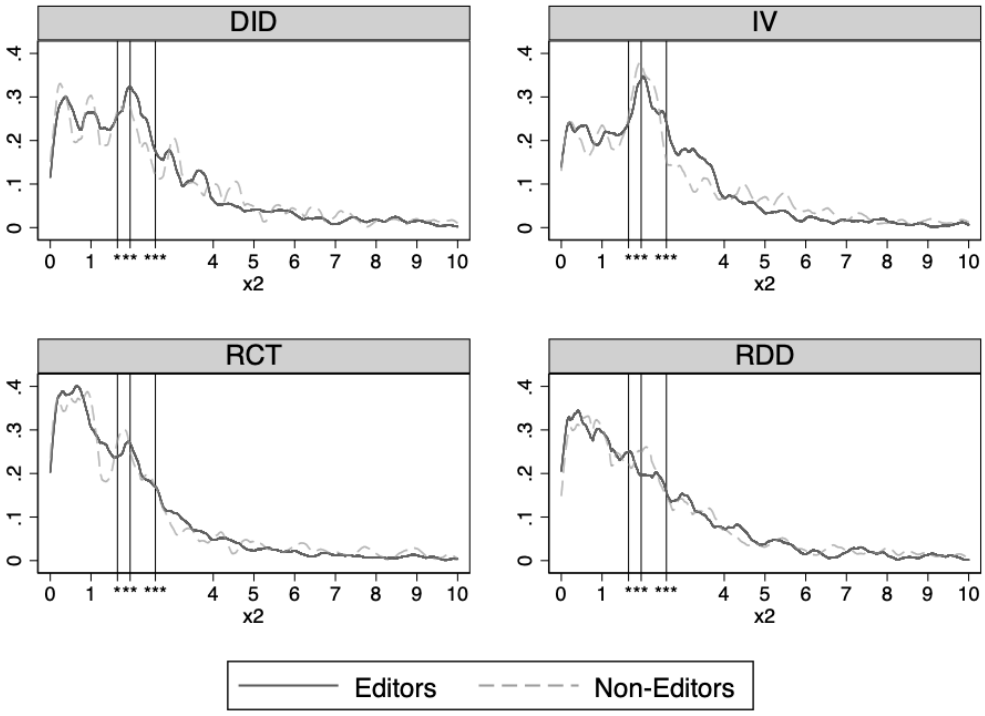
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Test statistics are partitioned by identification method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT) and regression discontinuity design (RDD). Lines in dark gray are for articles with at least one author who graduated from a top institution. Lines in light gray (dashes) are for articles with no author who graduated from a top institution. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure A7: z-Statistics by Method and Years of Experience



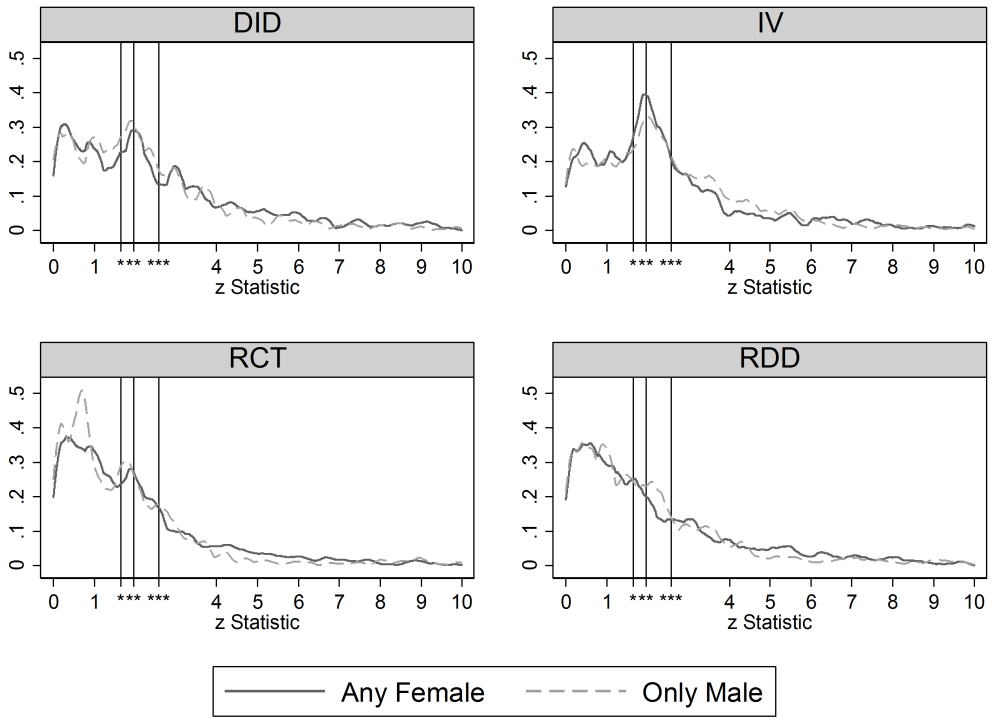
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Test statistics are partitioned by identification method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT) and regression discontinuity design (RDD). Lines in dark gray are for articles with authors having more than the median average years of experience (since PhD). Lines in light gray (dashes) are for articles with authors having less than the median average years of experience (since PhD). Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure A8: z-Statistics by Method and Editor



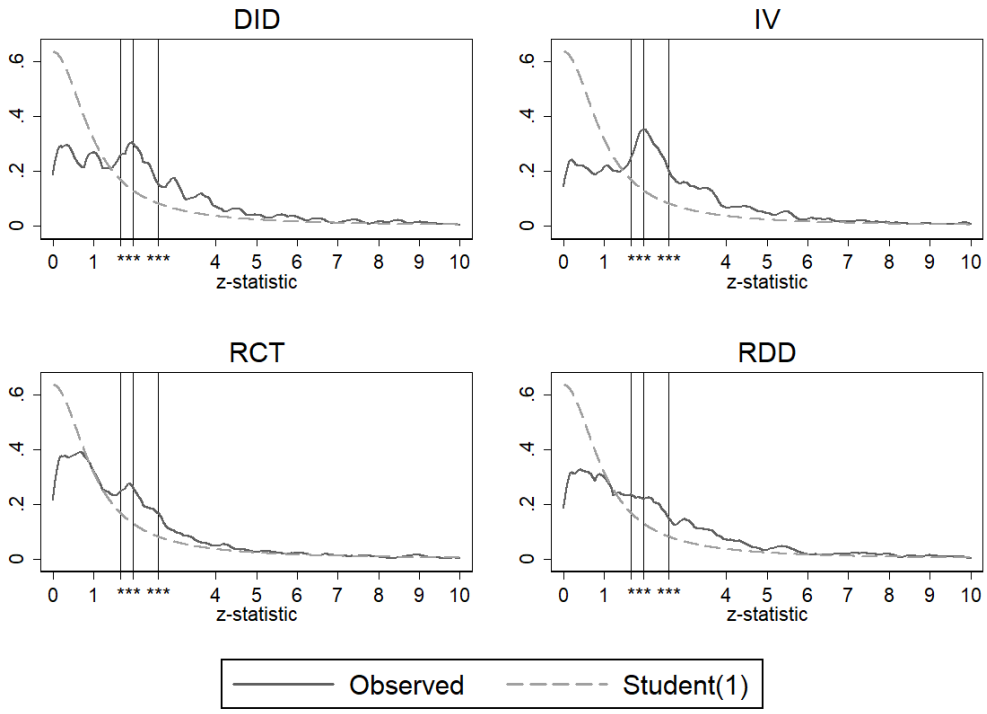
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Test statistics are partitioned by identification method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT) and regression discontinuity design (RDD). Lines in dark gray are for articles with at least one author being an editor of an economic journal. Lines in light gray (dashes) are for articles with no editors. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure A9: z-Statistics by Method and Authors' Gender



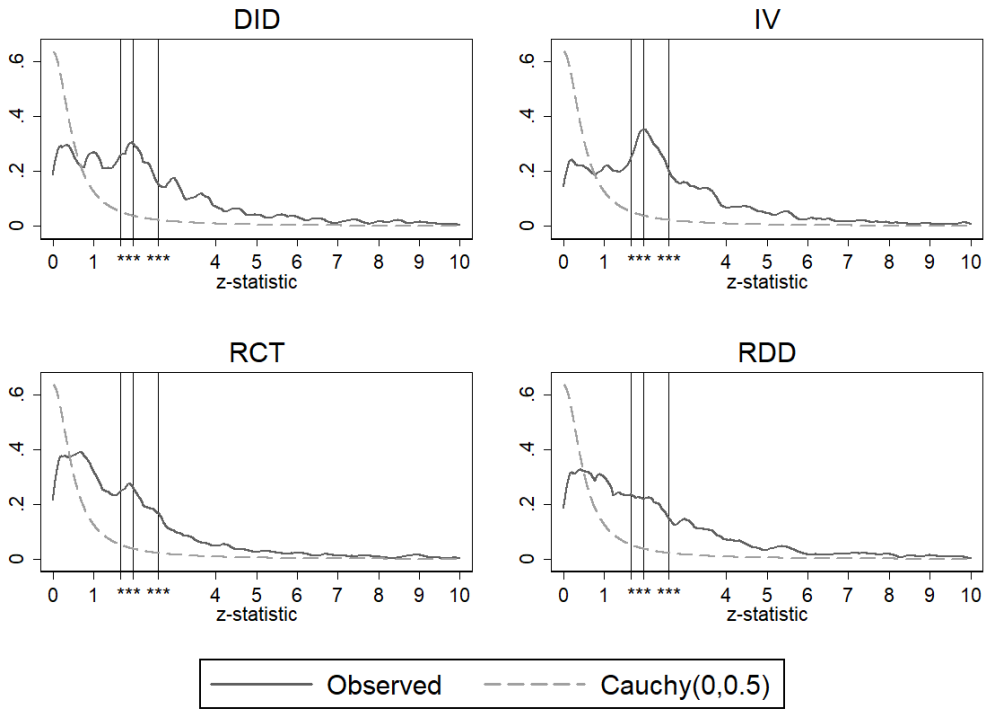
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. Test statistics are partitioned by identification method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT) and regression discontinuity design (RDD). Lines in dark gray are for articles with at least one female author. Lines in light gray are for articles with only male authors. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure A10: Observed Distributions and Student 1



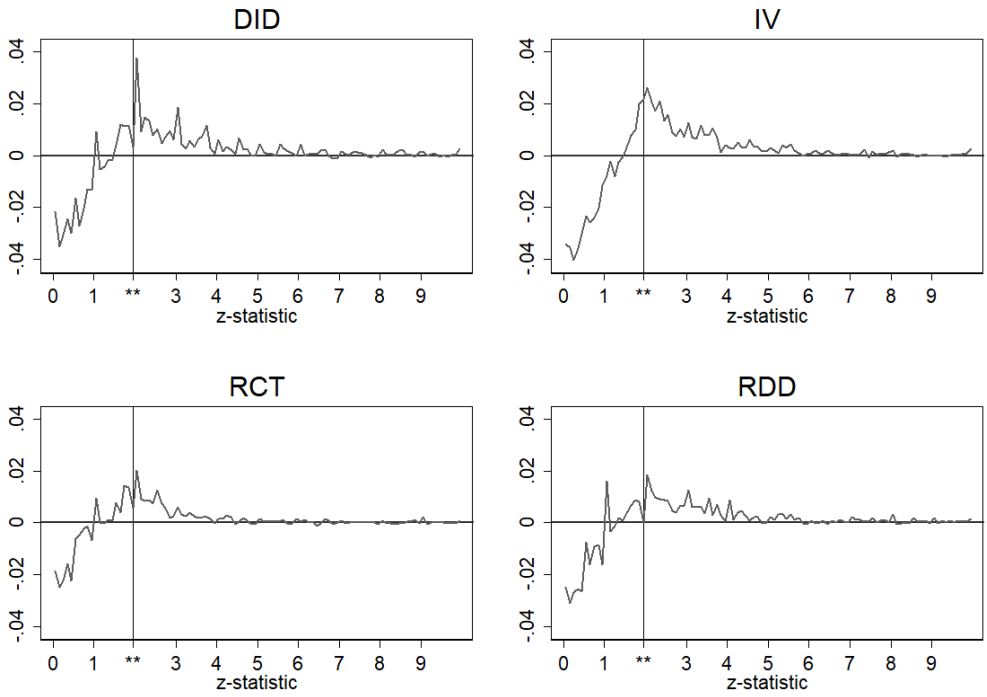
Notes: This figure displays the smoothed densities from Figure 2. The Student-t distribution with one degree of freedom is used as a reference distribution to detect excess (or missing) tests. Reference lines are displayed at the conventional two-tailed significance levels.

Figure A11: Observed Distributions and Cauchy 0,0.5



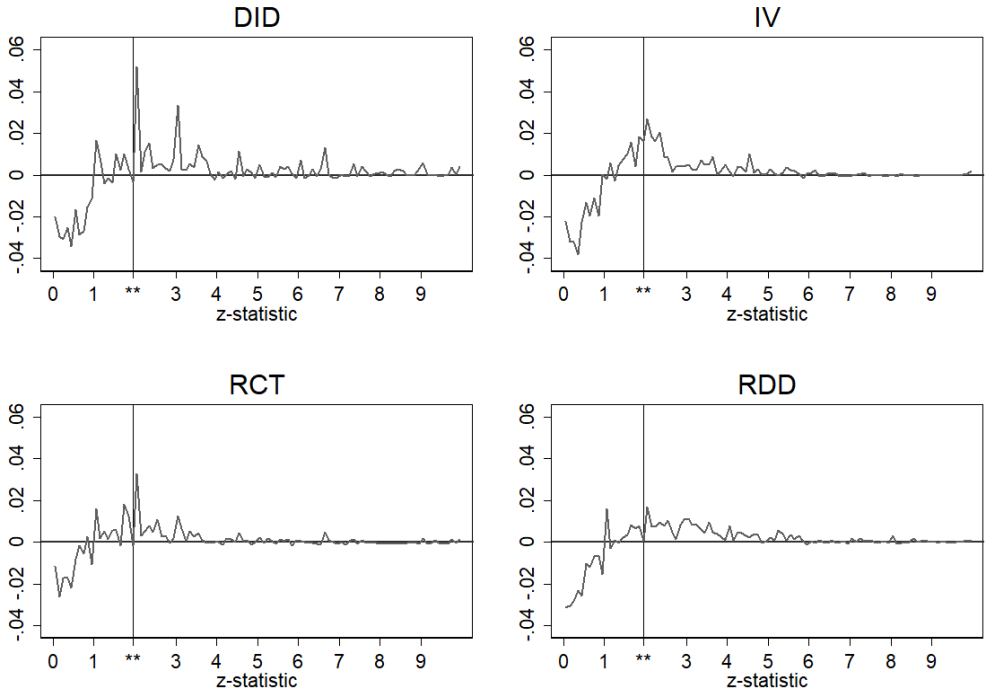
Notes: This figure displays the smoothed densities from Figure 2. The Cauchy(0,0.5) distribution is used as a reference distribution to detect excess (or missing) tests. Reference lines are displayed at the conventional two-tailed significance levels.

Figure A12: Differences Between Observed and Student-t Distributions



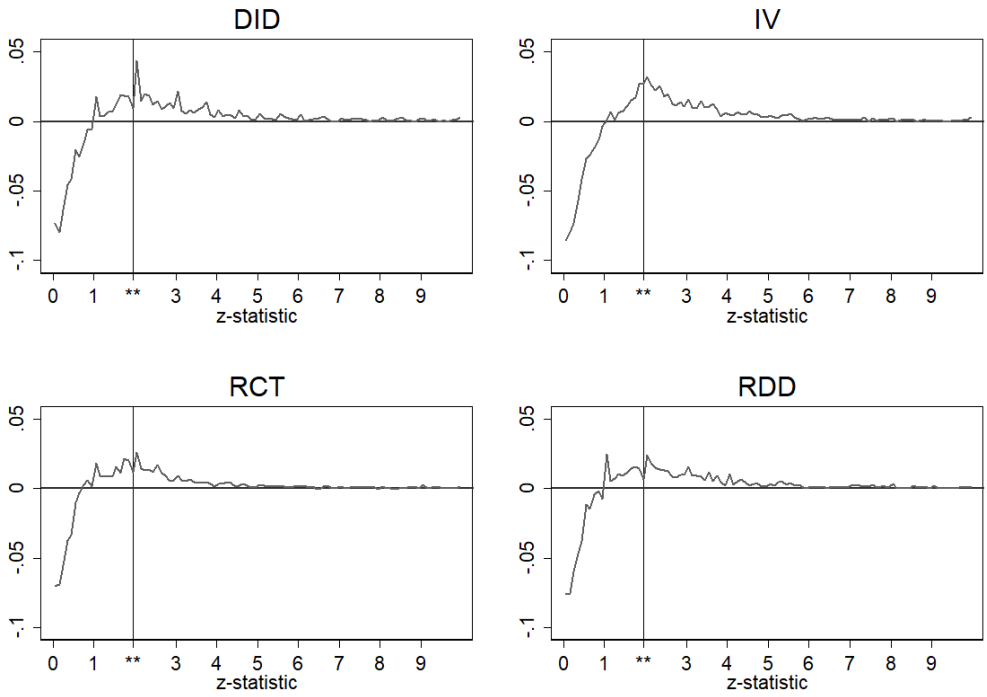
Notes: This figure displays the difference of the observed distribution to the input distribution. The input is a Student-t distribution with one degree of freedom. A negative number implies that there are 'missing' test statistics for that value of z whereas a positive number implies an excess of test statistics.

Figure A13: Differences Between Weighted Observed and Student-t Distributions



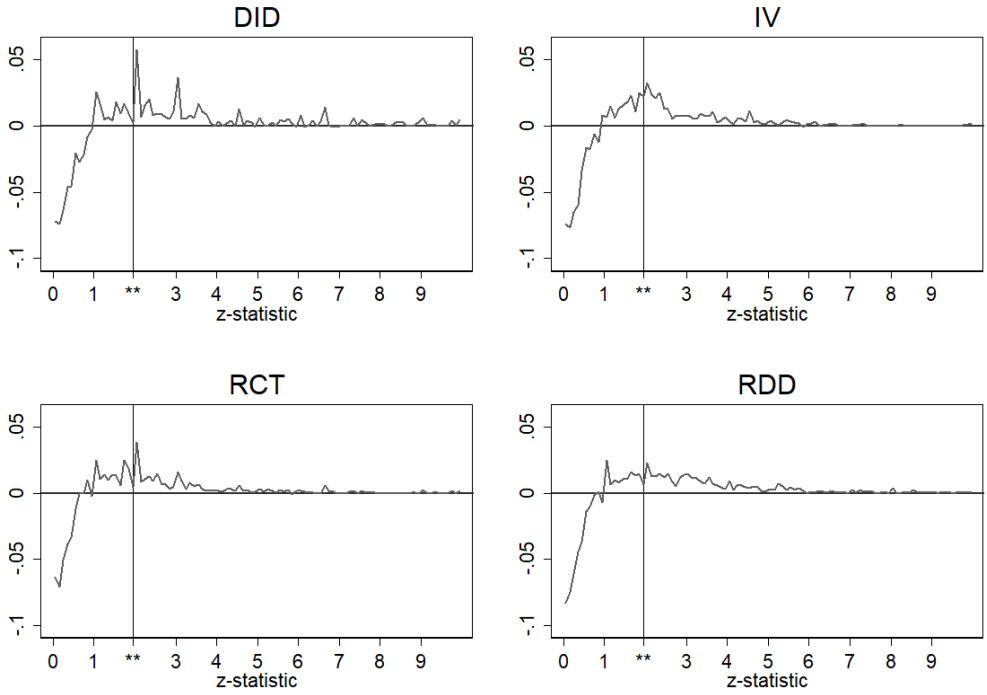
Notes: This figure displays the difference of the observed distribution to the input distribution. The input is a Student-t distribution with one degree of freedom. A negative number implies that there are 'missing' test statistics for that value of z whereas a positive number implies an excess of test statistics. We apply article × table weights to the observed test statistics.

Figure A14: Differences Between Observed and Cauchy Distributions



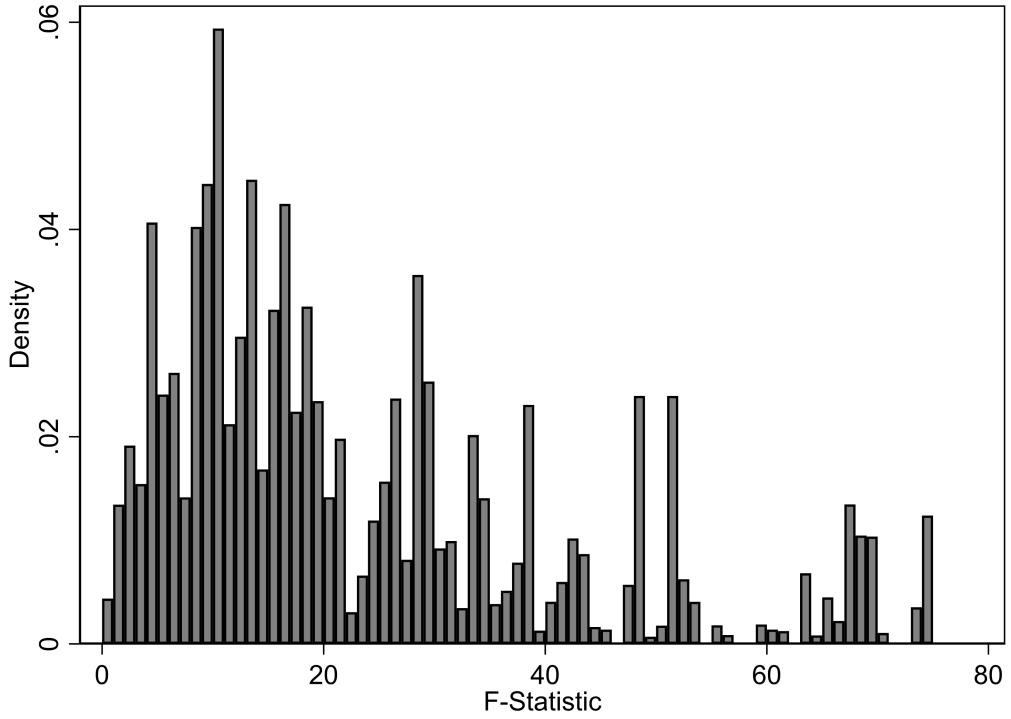
Notes: This figure displays the difference of the observed distribution to the input distribution. The input is a Cauchy(0,0.5) distribution. A negative number implies that there are 'missing' test statistics for that value of z whereas a positive number implies an excess of test statistics.

Figure A15: Differences Between Weighted Observed and Cauchy Distributions



Notes: This figure displays the difference of the observed distribution to the input distribution. The input is a Cauchy(0,0.5) distribution. A negative number implies that there are 'missing' test statistics for that value of z whereas a positive number implies an excess of test statistics. We apply article × table weights to the observed test statistics.

Figure A16: Instrumental Variable: First Stage F-Statistics



Notes: This figure displays an histogram of First Stage F-Statistics of instrumental variables for $F \in [0, 75]$.

Table A1: Articles Example

No	Pages	Text Flagged	Included as	Exclusion Notes
1	1-21	RCT		Meta analysis.
1	22-53	RCT	RCT	
1	54-89	RCT	RCT	
1	90-122	DID + IV + RCT	RCT	IV results "available upon request"
1	123-150	IV + RCT	IV + RCT	
1	151-182	RCT	RCT	
1	183-203	RCT	RCT	
2	1-34	RCT + RDD	RDD	Randomized trial mentioned in references only.
2	35-52	RCT		Randomized trial mentioned in literature review.
2	53-80	DID + RCT + RDD	DID	Discontinuity mentioned only in literature review.
2	81-108	IV + RCT	IV + RCT	
2	109-134	DID		Extended model.
2	135-174	RCT		Extended model.
2	175-206	DID+ RCT	DID	
2	207-232	IV + RCT	IV + RCT	
2	233-263	DID RD+ D		Extended model.
2	264-292	RCT		Plausible random assignment.
3	1-27	DID + RCT	DID	
3	28-50	IV + RCT + RDD	IV	Discontinuity mentioned only in references.
3	51-84	DID + RCT	RCT	
3	85-122	IV		Extended model.
3	123-146			Never text flagged.
3	147-177	IV + RCT	IV	
3	178-195	IV + RDD	RDD	Non-Standard IV
3	196-220	DID + RCT + RDD		Uses matching.
3	221-247			Never text flagged.
4	1-36	DID + IV + RCT	IV	
4	37-52	DID + IV + RCT		Non-standard IV.
4	53-75	DID + RCT	DID	
4	76-102	IV + RCT + RDD	IV	Discontinuity mentioned only in references.
4	103-135	IV		IV only in online appendix.
4	136-168	DID		Non-standard DID.
4	169-197			Never text flagged.
4	198-220	RCT + RDD	RDD	
4	221-253	DID + IV	DID + IV	

Notes: This table presents the 35 articles published in *American Economic Journal: Applied Economics* in 2015. We rely on the *American Economic Journal: Applied Economics* because it has all four methods, and is the first journal in our sample, alphabetically. Articles were text-searched using keywords, where * is a wildcard. For DID, "difference in difference*" "differences in difference*" "difference-in-difference*" and "differences-in-difference*" were used. For IV "instrumental variable*". For RCT "randomi*" and "control". For RDD "discontinuity".

Table A2: Significant at the 5% Level: Unweighted Estimates

	$Z > 1.96$ (1)	$Z > 1.96$ (2)	$Z > 1.96$ (3)	$Z > 1.96$ (4)	$Z > 1.96$ (5)	$Z > 1.96$ (6)
DID	0.132	0.132	0.127	0.120	0.128	0.133
	(0.046)	(0.046)	(0.050)	(0.050)	(0.047)	(0.047)
IV	0.171	0.171	0.169	0.164	0.158	0.168
	(0.046)	(0.046)	(0.047)	(0.045)	(0.044)	(0.044)
RDD	0.097	0.097	0.066	0.069	0.054	0.045
	(0.053)	(0.054)	(0.056)	(0.056)	(0.052)	(0.053)
Share PhD Top Institution						Y
Share Top Institution						Y
Share Female Author					Y	Y
Solo Authored					Y	Y
Experience					Y	Y
Experience Squared					Y	Y
Editor					Y	Y
ln(Citations)				Y	Y	Y
Reporting Method				Y	Y	Y
Table Ordering				Y	Y	Y
Journal FE			Y	Y	Y	Y
Top 5		Y	Y	Y	Y	Y
Observations	12,185	12,185	12,185	12,185	12,185	12,185
Pseudo R^2	0.012	0.012	0.032	0.037	0.041	0.043

Notes: This table reports marginal effects from probit regressions (Equation (1)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article.

Table A3: Covariates

	(1) $Z > 2.58$	(2) $Z > 1.96$	(3) $Z > 1.65$
Top 5	0.050	0.025	0.029
	(0.062)	(0.052)	(0.055)
t-Statistic	0.092	0.148	0.181
	(0.072)	(0.048)	(0.044)
p-Value	-0.049	-0.017	0.028
	(0.048)	(0.044)	(0.048)
Table Ordering	-0.017	-0.016	-0.011
	(0.006)	(0.006)	(0.006)
ln(Citations)	0.021	0.052	0.056
	(0.025)	(0.022)	(0.021)
Avg. Experience	-0.010	-0.015	-0.014
	(0.012)	(0.010)	(0.010)
Avg. Experience Squared	0.039	0.050	0.044
	(0.035)	(0.028)	(0.026)
Editor	-0.100	-0.071	-0.052
	(0.067)	(0.067)	(0.061)
Share Female Authors	0.031	0.022	0.034
	(0.056)	(0.054)	(0.053)
Single-Authored	-0.022	-0.003	-0.007
	(0.066)	(0.069)	(0.065)
Share Top Institution	-0.041	-0.043	-0.060
	(0.066)	(0.062)	(0.060)
Share Top PhD Institution	-0.013	-0.014	-0.022
	(0.049)	(0.044)	(0.044)
Observations	12,185	12,185	12,185
R-squared	0.023	0.030	0.035

Notes: This table reports marginal effects from probit regressions (Equation (1)). In column 1, the dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. In column 2, the dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. In column 3, the dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. Robust standard errors are in parentheses, clustered by article.

Table A4: Significant at the 10% Level

	$Z > 1.65$ (1)	$Z > 1.65$ (2)	$Z > 1.65$ (3)	$Z > 1.65$ (4)	$Z > 1.65$ (5)	$Z > 1.65$ (6)
DID	0.117	0.124	0.126	0.126	0.126	0.127
	(0.056)	(0.055)	(0.058)	(0.058)	(0.056)	(0.057)
IV	0.150	0.154	0.148	0.138	0.127	0.132
	(0.048)	(0.047)	(0.050)	(0.047)	(0.048)	(0.049)
RDD	0.027	0.027	0.016	0.008	-0.022	-0.035
	(0.061)	(0.062)	(0.054)	(0.054)	(0.054)	(0.056)
Share PhD Top Institution						Y
Share Top Institution						Y
Share Female Author					Y	Y
Solo-Authored					Y	Y
Experience					Y	Y
Experience Squared					Y	Y
Editor					Y	Y
ln(Citations)				Y	Y	Y
Reporting Method				Y	Y	Y
Table Ordering				Y	Y	Y
Journal FE			Y	Y	Y	Y
Top 5		Y	Y	Y	Y	Y
Observations	12,185	12,185	12,185	12,185	12,185	12,185
Pseudo R^2	0.014	0.014	0.040	0.047	0.054	0.057

Notes: This table reports marginal effects from probit regressions (Equation (1)). The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

Table A5: Significant at the 1% Level

	$Z > 2.58$ (1)	$Z > 2.58$ (2)	$Z > 2.58$ (3)	$Z > 2.58$ (4)	$Z > 2.58$ (5)	$Z > 2.58$ (6)
DID	0.118	0.126	0.112	0.100	0.097	0.099
	(0.075)	(0.077)	(0.082)	(0.080)	(0.076)	(0.075)
IV	0.053	0.057	0.051	0.047	0.030	0.033
	(0.059)	(0.059)	(0.060)	(0.056)	(0.056)	(0.056)
RDD	0.079	0.079	0.052	0.049	0.024	0.019
	(0.073)	(0.074)	(0.064)	(0.062)	(0.060)	(0.059)
Share PhD Top Institution						Y
Share Top Institution						Y
Share Female Author					Y	Y
Solo-Authored					Y	Y
Experience					Y	Y
Experience Squared					Y	Y
Editor					Y	Y
ln(Citations)				Y	Y	Y
Reporting Method				Y	Y	Y
Table Ordering				Y	Y	Y
Journal FE			Y	Y	Y	Y
Top 5		Y	Y	Y	Y	Y
Observations	12,185	12,185	12,179	12,179	12,179	12,179
Pseudo R^2	0.005	0.006	0.028	0.033	0.041	0.041

Notes: This table reports marginal effects from probit regressions (Equation (1)). The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

Table A6: Covariates, Caliper Test

	(1) $Z > 2.58$	(2) $Z > 1.96$	(3) $Z > 1.65$
Top 5	0.023	-0.068	-0.014
	(0.073)	(0.050)	(0.053)
t-Statistic	-0.084	0.024	0.146
	(0.084)	(0.077)	(0.056)
p-Value	-0.078	-0.150	-0.090
	(0.076)	(0.060)	(0.070)
Table Ordering	-0.006	-0.009	0.014
	(0.010)	(0.007)	(0.007)
ln(Citations)	-0.044	0.060	0.066
	(0.024)	(0.022)	(0.023)
Avg. Experience	0.000	-0.020	-0.030
	(0.014)	(0.014)	(0.014)
Avg. Experience Squared	0.004	0.069	0.101
	(0.052)	(0.061)	(0.058)
Editor	-0.072	-0.050	-0.109
	(0.064)	(0.060)	(0.054)
Share Female Authors	0.027	-0.048	-0.031
	(0.074)	(0.058)	(0.049)
Single-Authored	-0.005	-0.007	-0.113
	(0.075)	(0.064)	(0.063)
Share Top Institution	0.002	0.015	-0.072
	(0.074)	(0.051)	(0.048)
Share Top PhD Institution	0.008	0.039	-0.017
	(0.066)	(0.055)	(0.045)
Observations	2,102	2,922	2,921
R-squared	0.015	0.027	0.043

Notes: This table reports marginal effects from probit regressions (Equation (1)). In column 1, the dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. We restrict the sample to $z \in [2.08, 3.08]$. In column 2, the dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. We restrict the sample to $z \in [1.46, 2.46]$. In column 3, the dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. We restrict the sample to $z \in [1.15, 2.15]$. Robust standard errors are in parentheses, clustered by article.

Table A7: Caliper Test, Significant at the 10% Level

	(1) $Z > 1.65$	(2) $Z > 1.65$	(3) $Z > 1.65$	(4) $Z > 1.65$	(5) $Z > 1.65$	(6) $Z > 1.65$
DID	0.050	0.050	0.076	0.085	0.059	0.056
	(0.048)	(0.047)	(0.050)	(0.050)	(0.044)	(0.044)
IV	0.120	0.120	0.150	0.139	0.101	0.102
	(0.049)	(0.048)	(0.053)	(0.057)	(0.055)	(0.055)
RDD	-0.066	-0.066	-0.026	-0.036	-0.100	-0.113
	(0.050)	(0.050)	(0.053)	(0.055)	(0.054)	(0.054)
Share PhD Top Institution						Y
Share Top Institution						Y
Share Female Author					Y	Y
Solo-Authored					Y	Y
Experience					Y	Y
Experience Squared					Y	Y
Editor					Y	Y
ln(Citations)				Y	Y	Y
Reporting Method				Y	Y	Y
Table Ordering				Y	Y	Y
Journal FE			Y	Y	Y	Y
Top 5		Y	Y	Y	Y	Y
Observations	2,921	2,921	2,921	2,921	2,921	2,921
Pseudo R-squared	0.016	0.016	0.034	0.045	0.059	0.060
Window Width			$Z > 1.15$ & $Z < 2.15$			

Notes: This table reports marginal effects from probit regressions (Equation (1)). The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. We restrict the sample to $z \in [1.15, 2.15]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

Table A8: Caliper Test, Significant at the 1% Level

	(1) $Z > 2.58$	(2) $Z > 2.58$	(3) $Z > 2.58$	(4) $Z > 2.58$	(5) $Z > 2.58$	(6) $Z > 2.58$
DID	-0.002	-0.008	-0.035	-0.060	-0.066	-0.066
	(0.071)	(0.073)	(0.090)	(0.081)	(0.082)	(0.082)
IV	-0.034	-0.037	-0.054	-0.058	-0.075	-0.082
	(0.070)	(0.072)	(0.078)	(0.075)	(0.075)	(0.076)
RDD	0.028	0.026	-0.005	-0.003	-0.030	-0.028
	(0.072)	(0.074)	(0.076)	(0.073)	(0.067)	(0.067)
Share PhD Top Institution						Y
Share Top Institution						Y
Share Female Author					Y	Y
Solo-Authored					Y	Y
Experience					Y	Y
Experience Squared					Y	Y
Editor					Y	Y
ln(Citations)				Y	Y	Y
Reporting Method				Y	Y	Y
Table Ordering				Y	Y	Y
Journal FE			Y	Y	Y	Y
Top 5		Y	Y	Y	Y	Y
Observations	2,102	2,102	2,102	2,102	2,102	2,102
Pseudo R-squared	0.002	0.002	0.024	0.033	0.037	0.037
Window Width			$Z > 2.08$ & $Z < 3.08$			

Notes: This table reports marginal effects from probit regressions (Equation (1)). The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. We restrict the sample to $z \in [2.08, 3.08]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

Table A9: Caliper Test, Significant at the 5% Level: Windows

	(1) $Z > 1.96$	(2) $Z > 1.96$	(3) $Z > 1.96$	(4) $Z > 1.96$	(5) $Z > 1.96$
DID	0.131	0.136	0.148	0.143	0.148
	(0.046)	(0.048)	(0.050)	(0.054)	(0.061)
IV	0.138	0.138	0.133	0.099	0.090
	(0.047)	(0.048)	(0.050)	(0.055)	(0.056)
RDD	0.008	0.017	0.010	0.022	0.012
	(0.046)	(0.049)	(0.052)	(0.052)	(0.053)
Share Female Author	Y	Y	Y	Y	Y
Solo-Authored	Y	Y	Y	Y	Y
Experience	Y	Y	Y	Y	Y
Experience Squared	Y	Y	Y	Y	Y
Editor	Y	Y	Y	Y	Y
ln(Citations)	Y	Y	Y	Y	Y
Reporting Method	Y	Y	Y	Y	Y
Table Ordering	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y
Top 5	Y	Y	Y	Y	Y
Observations	3,392	3,174	2,922	2,632	2,401
Pseudo R^2	0.032	0.035	0.042	0.046	0.045
Window Width $z \in$	$[1.36, 2.56]$	$[1.41, 2.51]$	$[1.46, 2.46]$	$[1.51, 2.41]$	$[1.56, 2.36]$

Notes: This table reports marginal effects from probit regressions (Equation (1)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. In column 1, we restrict the sample to $z \in [1.36, 2.56]$. Column 2 restricts the sample to $z \in [1.41, 2.51]$. In column 3, we restrict the sample to $z \in [1.46, 2.46]$. Column 4 restricts the sample to $z \in [1.51, 2.41]$. In column 5, we restrict the sample to $z \in [1.56, 2.36]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

Table A10: Caliper Test, Significant at the 10% Level: Windows

	(1) $Z > 1.65$	(2) $Z > 1.65$	(3) $Z > 1.65$	(4) $Z > 1.65$	(5) $Z > 1.65$
DID	0.081	0.060	0.059	0.080	0.082
	(0.043)	(0.044)	(0.044)	(0.044)	(0.047)
IV	0.109	0.090	0.101	0.103	0.105
	(0.053)	(0.055)	(0.055)	(0.054)	(0.058)
RDD	-0.079	-0.097	-0.100	-0.088	-0.080
	(0.051)	(0.055)	(0.054)	(0.052)	(0.055)
Share Female Author	Y	Y	Y	Y	Y
Solo-Authored	Y	Y	Y	Y	Y
Experience	Y	Y	Y	Y	Y
Experience Squared	Y	Y	Y	Y	Y
Editor	Y	Y	Y	Y	Y
ln(Citations)	Y	Y	Y	Y	Y
Reporting Method	Y	Y	Y	Y	Y
Table Ordering	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y
Top 5	Y	Y	Y	Y	Y
Observations	3,447	3,205	2,921	2,641	2,357
Pseudo R^2	0.047	0.049	0.059	0.063	0.060
Window Width $z \in$	$[1.05, 2.25]$	$[1.10, 2.20]$	$[1.15, 2.15]$	$[1.20, 2.10]$	$[1.25, 2.05]$

Notes: This table reports marginal effects from probit regressions (Equation (1)). The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. In column 1, we restrict the sample to $z \in [1.05, 2.25]$. Column 2 restricts the sample to $z \in [1.10, 2.20]$. In column 3, we restrict the sample to $z \in [1.15, 2.15]$. Column 4 restricts the sample to $z \in [1.20, 2.10]$. In column 5, we restrict the sample to $z \in [1.25, 2.05]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

Table A11: Significant at the 5% Level: Logit

	$Z > 1.96$ (1)	$Z > 1.96$ (2)	$Z > 1.96$ (3)	$Z > 1.96$ (4)	$Z > 1.96$ (5)	$Z > 1.96$ (6)
DID	0.150	0.156	0.150	0.144	0.144	0.143
	(0.060)	(0.061)	(0.066)	(0.066)	(0.061)	(0.061)
IV	0.139	0.142	0.139	0.134	0.119	0.122
	(0.050)	(0.049)	(0.053)	(0.049)	(0.049)	(0.049)
RDD	0.055	0.055	0.039	0.036	-0.001	-0.008
	(0.063)	(0.063)	(0.056)	(0.054)	(0.053)	(0.054)
Share PhD Top Institution						Y
Share Top Institution						Y
Share Female Author					Y	Y
Solo-Authored					Y	Y
Experience					Y	Y
Experience Squared					Y	Y
Editor					Y	Y
ln(Citations)				Y	Y	Y
Reporting Method				Y	Y	Y
Table Ordering				Y	Y	Y
Journal FE			Y	Y	Y	Y
Top 5		Y	Y	Y	Y	Y
Observations	12,185	12,185	12,185	12,185	12,185	12,185
Pseudo R^2	0.011	0.011	0.032	0.039	0.046	0.047

Notes: This table reports marginal effects from logit regressions (Equation (1)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

Table A12: Caliper Test, Significant at the 5% Level: Logit

	$Z > 1.96$ (1)	$Z > 1.96$ (2)	$Z > 1.96$ (3)	$Z > 1.96$ (4)	$Z > 1.96$ (5)	$Z > 1.96$ (6)
DID	0.158	0.154	0.158	0.155	0.150	0.151
	(0.048)	(0.048)	(0.055)	(0.055)	(0.050)	(0.050)
IV	0.106	0.105	0.125	0.154	0.136	0.131
	(0.046)	(0.045)	(0.050)	(0.052)	(0.050)	(0.050)
RDD	0.003	0.004	0.027	0.052	0.010	0.012
	(0.038)	(0.036)	(0.049)	(0.050)	(0.053)	(0.054)
Share PhD Top Institution						Y
Share Top Institution						Y
Share Female Author					Y	Y
Solo-Authored					Y	Y
Experience					Y	Y
Experience Squared					Y	Y
Editor					Y	Y
ln(Citations)				Y	Y	Y
Reporting Method				Y	Y	Y
Table Ordering				Y	Y	Y
Journal FE			Y	Y	Y	Y
Top 5		Y	Y	Y	Y	Y
Observations	2,922	2,922	2,922	2,922	2,922	2,922
Pseudo R^2	0.012	0.012	0.030	0.039	0.042	0.042
Window Width			$Z > 1.46$ & $Z < 2.46$			

Notes: This table reports marginal effects from logit regressions (Equation (1)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

Table A13: Significant at the 1%, 5% and 10% Level: Coding/Data Collection

	(1) $Z > 2.58$	(2) $Z > 1.96$	(3) $Z > 1.65$	(4) $Z > 2.58$	(5) $Z > 1.96$	(6) $Z > 1.65$
DID	0.108	0.152	0.133	0.090	0.125	0.114
	(0.075)	(0.061)	(0.056)	(0.082)	(0.066)	(0.056)
IV	0.029	0.118	0.129	0.055	0.117	0.134
	(0.056)	(0.049)	(0.049)	(0.063)	(0.053)	(0.052)
RDD	0.016	-0.013	-0.038	0.056	0.010	-0.013
	(0.059)	(0.053)	(0.055)	(0.068)	(0.061)	(0.059)
Share PhD Top Institution	Y	Y	Y	Y	Y	Y
Share Top Institution	Y	Y	Y	Y	Y	Y
Share Female Author	Y	Y	Y	Y	Y	Y
Solo-Authored	Y	Y	Y	Y	Y	Y
Experience	Y	Y	Y	Y	Y	Y
Experience Squared	Y	Y	Y	Y	Y	Y
Editor	Y	Y	Y	Y	Y	Y
ln(Citations)	Y	Y	Y	Y	Y	Y
Reporting Method	Y	Y	Y	Y	Y	Y
Table Ordering	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y
Top 5	Y	Y	Y	Y	Y	Y
Observations	12,020	12,026	12,026	10,484	10,490	10,490
Pseudo R^2	0.043	0.048	0.057	0.050	0.052	0.060

Notes: This table reports marginal effects from logit regressions (Equation (1)). The dependent variables are dummies for whether the test statistic is significant at the 1, 5 and 10 percent level, respectively. Columns 1–3 drop the articles for which we could not easily reach an agreement on which tests to select. In columns 4–6, we exclude journal articles relying on multiple methods. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

Table A14: Caliper Test, Significant at the 1%, 5% and 10% Level: Coding/Data Collection

	(1) $Z > 2.58$	(2) $Z > 1.96$	(3) $Z > 1.65$	(4) $Z > 2.58$	(5) $Z > 1.96$	(6) $Z > 1.65$
DID	-0.062	0.155	0.060	-0.110	0.125	0.048
	(0.083)	(0.051)	(0.044)	(0.079)	(0.060)	(0.054)
IV	-0.086	0.132	0.102	-0.038	0.108	0.094
	(0.077)	(0.050)	(0.056)	(0.076)	(0.054)	(0.065)
RDD	-0.033	0.015	-0.112	0.032	0.014	-0.120
	(0.067)	(0.053)	(0.054)	(0.074)	(0.056)	(0.061)
Share PhD Top Institution	Y	Y	Y	Y	Y	Y
Share Top Institution	Y	Y	Y	Y	Y	Y
Share Female Author	Y	Y	Y	Y	Y	Y
Solo-Authored	Y	Y	Y	Y	Y	Y
Experience	Y	Y	Y	Y	Y	Y
Experience Squared	Y	Y	Y	Y	Y	Y
Editor	Y	Y	Y	Y	Y	Y
ln(Citations)	Y	Y	Y	Y	Y	Y
Reporting Method	Y	Y	Y	Y	Y	Y
Table Ordering	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y
Top 5	Y	Y	Y	Y	Y	Y
Observations	2,082	2,897	2,892	1,763	2,501	2,524
Pseudo R^2	0.037	0.043	0.059	0.042	0.040	0.062
Window Width $z \in$	$[2.08, 3.08]$	$[1.46, 2.46]$	$[1.15, 2.15]$	$[2.08, 3.08]$	$[1.46, 2.46]$	$[1.15, 2.15]$

Notes: This table reports marginal effects from logit regressions (Equation (1)). The dependent variables are dummies for whether the test statistic is significant at the 1, 5 and 10 percent level, respectively. Columns 1–3 drop the articles for which we could not easily reach an agreement on which tests to select. In columns 4–6, we exclude journal articles relying on multiple methods. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.

Table A15: Robustness Check: Omission of Economic Sub-Fields

	$Z > 1.96$ (1)	$Z > 1.96$ (2)	$Z > 1.96$ (3)
Field Omitted	Top 5	Finance	Macroeconomics
DID	0.201	0.061	0.143
	(0.060)	(0.068)	(0.061)
IV	0.116	0.114	0.121
	(0.062)	(0.055)	(0.049)
RDD	0.045	-0.007	-0.011
	(0.062)	(0.058)	(0.054)
Observations	9,828	10,264	12,142
Field Omitted	General Interest	Development	Experimental
DID	0.146	0.144	0.146
	(0.070)	(0.062)	(0.061)
IV	0.105	0.121	0.126
	(0.057)	(0.047)	(0.049)
RDD	-0.017	-0.026	-0.007
	(0.062)	(0.053)	(0.054)
Observations	8,504	11,295	12,156
Field Omitted	Labor	Public	Urban
DID	0.167	0.142	0.138
	(0.061)	(0.069)	(0.062)
IV	0.154	0.119	0.106
	(0.048)	(0.050)	(0.049)
RDD	0.016	-0.045	-0.003
	(0.053)	(0.058)	(0.053)
Observations	11,756	9,992	11,784
Articles' Characteristics	Y	Y	Y
Autors' Characteristics	Y	Y	Y
Journal FE	Y	Y	Y
Top 5	Y	Y	Y

Notes: This table reports marginal effects from probit regressions (Equation (1)). We omit an economic sub-field in each regression. The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same table multiplied by the inverse of the number of tables in the article to weight observations.