CAHIER DE RECHERCHE #1809E
Département de science économique
Faculté des sciences sociales
Université d'Ottawa

WORKING PAPER #1809E
Department of Economics
Faculty of Social Sciences
University of Ottawa

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

Abel Brodeur,† Nikolai Cook,‡ Anthony Heyes§

August 2018

* We are grateful to Jason Garred and seminar participants at Carleton University and the University of Ottawa for useful remarks and encouragement. We thank Richard Beard for research assistance. Errors are ours.
† Department of Economics, University of Ottawa; 120 University Private, Ottawa, Ontario, Canada, K1N 6N5; e-mail: abrodeur@uottawa.ca
‡ Department of Economics, University of Ottawa; 120 University Private, Ottawa, Ontario, Canada, K1N 6N5; e-mail: ncook@uottawa.ca
§ Department of Economics, University of Ottawa; 120 University Private, Ottawa, Ontario, Canada, K1N 6N5, and University of Sussex; e-mail: anthony.heyes@uottawa.ca

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 13,440 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. Almost 25% of claims of marginally significant results in IV papers are misleading.

Key words: Research methods, causal inference, p-curves, p-hacking, publication bias

JEL Classification: A11, B41, C13, C44.

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?

Evidence of p-hacking and publication bias in economics and other disciplines is by now voluminous (for examples Vivalt (2017); Casey et al. (2012); Gerber and Malhotra (2008a); Gerber and Malhotra (2008b); Havránek (2015); Henry (2009); Ioannidis (2005); Leamer (1983); Leamer and Leonard (1983) Ridley et al. (2007) Simmons et al. (2011); Stanley (2008)). Publication bias may reflect that the peer review process makes the likelihood that a particular result is published sensitive to the statistical significance (or otherwise) attached to it. 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.

In recent years, empirical economics has experienced a "credibility revolution" with a shift in focus to causal inference. As a result, experimental and quasi-experimental methods have become the norm (Angrist and Pischke, 2010). The four methods in common use are difference-in-differences (DID), instrumental variables (IV), randomized control trials (RCT), and the regression discontinuity design (RDD). Our interest is in a different sort of credibility, the extent to which claims made about statistical significance in papers using each of these methods are reliable. Harvesting 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.

We employ three approaches in our analysis.

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 13 to 16% more likely to report test statistics that are statistically significant at the 5% level. 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 field of study and to 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). 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 DID and (in particular) IV papers looking the most heavily biased.

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. Our results are broadly consistent with what we believe to be the 'folk wisdom' in the economics profession, namely that RCT and RDD are comparatively difficult to

manipulate, while IV and DID methods give researchers more discretion or 'wiggle room' (e.g., Young (2018)). These insights should be of interest to any reader of empirical research in economics.

1 Data and Methods

We collect the universe of articles published by 25 top journals in economics during the 2015 calendar year. Appendix Table A1 provides the complete list of journals included in the analysis. We selected top journals as ranked using RePEc's Simple Impact Factor.[1] 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.

To identify method sub-samples from this universe, we search the title and text body of each article for specific terms. We search "difference-in-differences" or "differences-in-differences", with and without the dashes, for DID, "instrumental variable" for IV, "random" and "trial" for RCT and "regression discontinuity" for RDD articles. Not all articles that contained the search word were ultimately used. We manually excluded articles that contained the search term but did not use one of the four methods. For example, some articles contained the search term as a title of another paper in their bibliographies. Furthermore, articles are excluded if they use a sub-method that affords additional researcher freedoms, e.g., matching in the case of DID or an instrumental variable designed to conduct a fuzzy RDD. Articles that restrict researcher freedoms such as structural equation modeling articles are also removed. We ultimately collect statistics from 308 articles.

Within article, tables are excluded if they present summary statistics. Our goal is to collect only coefficients of interest, or main results. For this reason, appendices, robustness checks, and placebo tests are similarly excluded from collection. 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.

[1] 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`.

Within tables, only coefficients of interest are collected - we exclude any obvious regression controls or constant terms. Coefficients drawn from multiple specifications are collected. All reported decimal places are collected. 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. We ultimately collect a total of 13,440 test statistics. On average, we collect 34 estimates per article for DID and 24 estimates per IV article. RCT and RDD offer 60 and 92 coefficients per article. In our analyses we include article weights to prevent articles with more tests from having a disproportionate effect. Appendix Table A1 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 23%, 21%, 26% and 30% respectively.

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 deviation 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.

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 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 a plausible counterfactual distribution.

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

$$Significant_{ijf} = \alpha + \beta_j + \delta_f + \gamma DID_{ijf} + \lambda IV_{ijf} + \phi RDD_{ijf} + \varepsilon_{ijf} \tag{1}$$

where $Significant_{ijf}$ is an indicator variable that estimate i is statistically significant in journal j and field f. We also include indicators for individual journals. Our results hold within-journal. We also include indicators for reporting methods, i.e., p-values or t-statistics. For the caliper test, we restrict the sample to a narrow band around a statistical significance threshold. We report marginal effects from probit regression throughout. Standard errors are clustered at article level.

2 Results

2.1 Descriptive Analysis

Figure 1 displays barcharts of z-statistics for each of the four methods. 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 conventional significance levels. 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 2, 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 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.

2.2 Probit Estimates: Whole Sample

Table 1 presents estimates of Equation 1 where the dependent variable indicates whether a test statistic is statistically significant at the 5% level. (Appendix Tables A3 and A4 replicate Table 1 for the other conventional significance levels.) 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.[2] In the most parsimonious specification, we find that a DID estimate is 13% more likely to be statistically significant than a RCT estimate. An IV estimate is 18% more likely to be statistically significant than a RCT estimate. Our estimates are statistically significant at the 1% 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 covariates and fixed effects. In column 2, we include an indicator for whether the estimate is in an article published in a top 5 journal.[3] In column 3, we include individual journal fixed effects. In column 4, we add indicators for reporting t-statistic or p-values directly. Our estimates are statistically

[2]Clustering by journal yields very similar conclusions.

[3]Top 5 journals in economics are the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics* and *Review of Economic Studies*. We do not find that estimates in the Top 5 are more or less likely to be statistically significant than estimates outside of the Top 5, conditional (or not) on method used. See Appendix Table A5.

significant across specifications and range from 13.0% to 13.4% for DID and from 17.7% to 18.6% for IV. On the other hand, the RDD estimates are consistently small, positive and statistically insignificant.

In column 5, we show that our results are robust to an alternative but informative weighting scheme. We weight test coefficients using the inverse of the number of tests presented in the same article. Our estimates are slightly larger for DID and slightly smaller for IV. These results confirm our conclusions are not driven by articles with many or few estimates.

Overall, these findings provide evidence that within top journals in economics, the likelihood of a null hypothesis being rejected depends upon the underlying research method. We test in the next subsection whether these patterns hold when we restrict the sample to marginally significant and insignificant tests.

2.3 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 2 reports the estimates for the 5 percent significance level. The structure of the table is the same as in Table 1. The only difference is that we restrict the sample to $z \in [1.46, 2.46]$. Our sample size decreases to 3,171 observations.

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 6.6% to 9.3% (RCT is the baseline). On the other hand, we find slightly weaker evidence that DID articles are more likely to report marginally significant tests. The estimates are significant at conventional levels only in columns 4 and 5.

Similarly, we present estimates for the other conventional levels in Appendix Tables A6

and A7. We also rely on a window of ±0.5 for the other two significance levels.[4] We confirm that IV articles are significantly more likely to report marginally significant tests at the 10 percent level than RCT and RDD. But we do not find any evidence that the extent of p-hacking varies by methods for the 1 percent significance threshold. All our estimates are small and statistically insignificant.

To sum up, 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 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.

2.4 Excess Test Statistics

For our third exercise we seek to quantify an excess (or dearth) of test statistics in particular ranges. To do this we take three steps.

The first step is to generate a smoothed density from the observed coefficients. These correspond directly with the overlaid densities in Figure 1.

Second, we construct the difference between each of the observed distributions and a hypothetical distribution. Intuitively, the 'natural' distribution of tests is likely to have a decreasing pattern over the whole interval. Appropriate counterfactual distributions - what the distribution of z-statistics should look like absent any manipulation or publication bias - have been explored empirically by performing large numbers of random tests on large datasets (Brodeur et al., 2016). Following Brodeur et al. (2016) we use the Student 1 distribution as the hypothetical sampling distribution of coefficients. (See Appendix Figure A1.) In a robustness check, we also use Cauchy distributions, e.g., Cauchy of parameter 0.5.[5] The "excess" test statistics for each method is presented graphically in Appendix

[4] Our findings are robust to other windows such as ±0.45, ±0.55, ±0.60 or ±0.65. Estimates available upon request.

[5] Note that our results are robust to the use of distributions derived empirically. These are not shown for space consideration. We rely on the same datasets as in Brodeur et al. (2016). In their study, they randomly draw variables from four economic datasets and run millions of regressions between these variables. The

Figure A2. For each of the methods, there is a dearth of insignificant coefficients and an excess of significant coefficients. Visually, RDD performs best relative to the other methods, followed by RCT and DID. IV performs the worst.

Third, we sum the differences over our interval of interest: $z \in [1, 3]$. These sums correspond to the areas below the curves in Figure A2 and are presented in Table 3. Our findings suggest that, respectively, 9.8% and 12.6% of tests in $z \in [1.00, 1.65]$ are misallocated for RDD and RCT. In contrast, 23.8% of IV tests are misallocated, with most of the misallocated tests in the interval $[1.96, 3]$. Misallocation is also quite large for DID with approximately 18% of tests misallocated.

3 Discussion

Concerns about p-hacking and publication bias have emerged in recent years (Brodeur et al. (2016)). Our results suggest that published studies using DID and IV are more p-hacked than RCT and RDD. Our estimates suggest that the extent of misallocation is the greatest for IV with almost 25% of marginally rejected tests misallocated and the smallest for RCT and RDD. We believe this to be roughly consistent with folk wisdom within the profession, which often regards RCT as a gold standard, and treat papers employing IV methods with the greatest 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. While this may partly explain the large (small) number of RCT and RDD (IV and DID) studies with high (low) p-values, we think it cannot explain the extent of p-hacking by method. In other words, different research questions may lead to different rejection rates, but not only for estimates that are marginally (in)significant.

This paper contributes to a discussion of the trustworthiness or dependability of empirical claims made by economics researchers (Leamer (1983)). Aided by access to better data and advancements in theoretical econometrics, design-based research methods are credited as

interested reader is directed to that paper to understand the case for the counterfactual distributions that we adopt for current purposes.

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.[6]

[6]The 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)).

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.

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/.

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).

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.

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.

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.

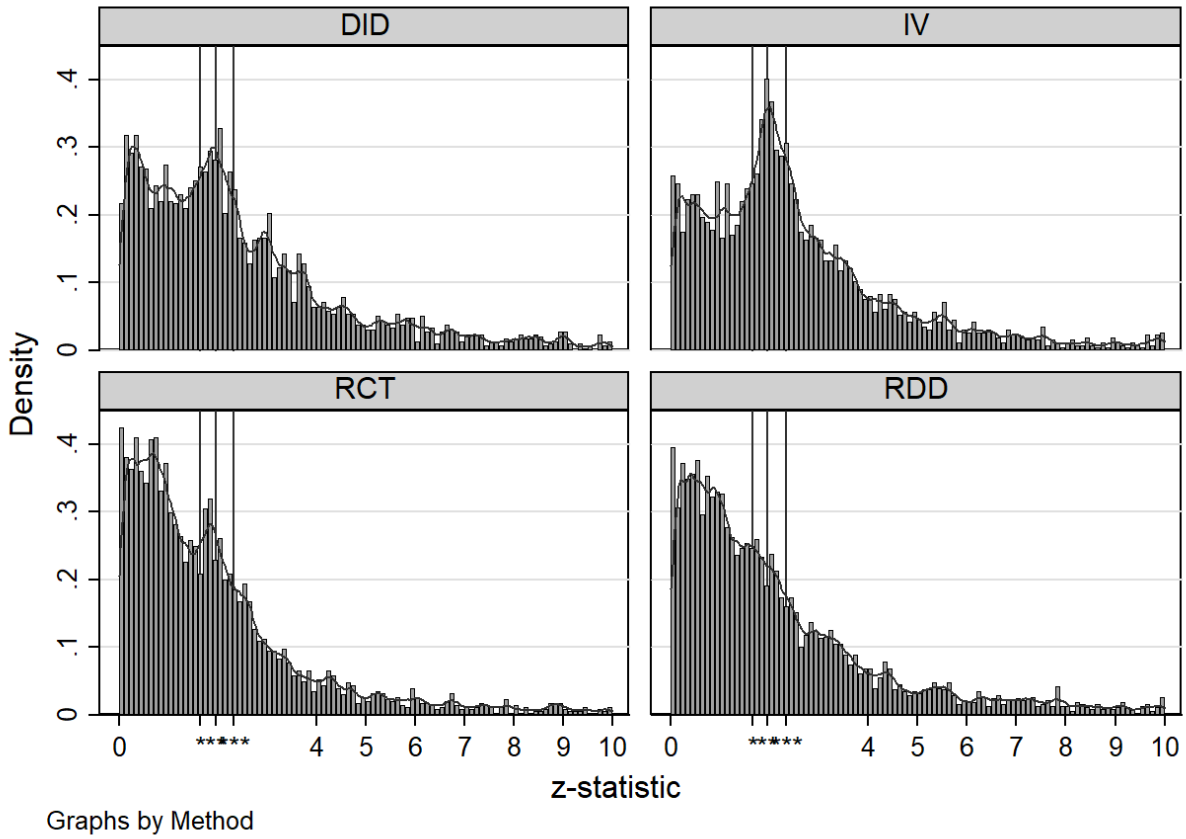
Stanley, T. D. (2008). Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection. *Oxford Bulletin of Economics and Statistics*, 70(1):103–127.

Vivalt, E. (2017). The Trajectory of Specification Searching and Publication Bias Across Methods and Disciplines. mimeo: Australian National University.

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

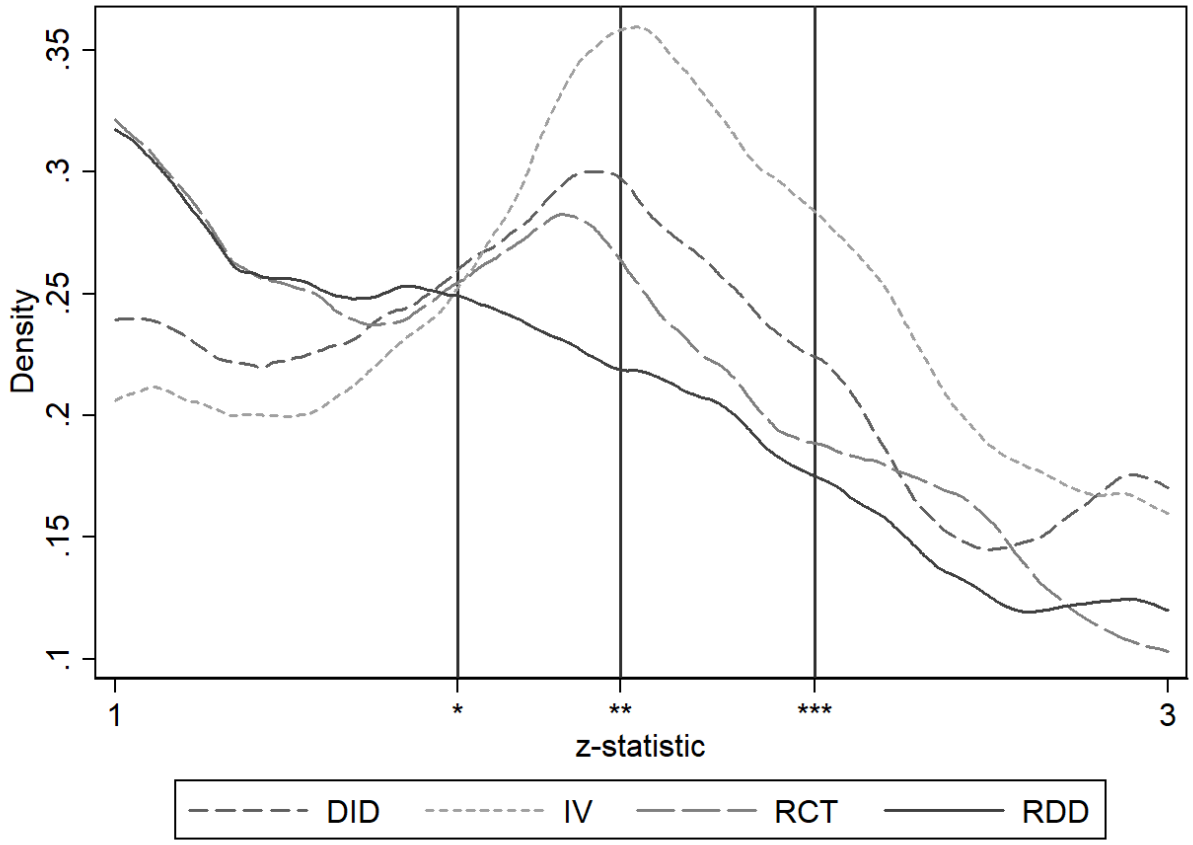
4 Figures

Figure 1: z-Statistics by Method



This figure displays histograms of test-coefficients for $z \in [0, 10]$. Coefficients 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 conventional significance levels. We have also superimposed an Epanechnikov kernel. Test statistics have been de-rounded.

Figure 2: z-Curves by Method



This figure displays the smoothed densities (Epanechnikov) from Figure 1 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 conventional significance levels.

5 Tables

Table 1: Significant at the 5% Level

	(1) $Z > 1.96$	(2) $Z > 1.96$	(3) $Z > 1.96$	(4) $Z > 1.96$	(5) $Z > 1.96$
DID	0.132	0.133	0.134	0.130	0.163
	(0.044)	(0.044)	(0.046)	(0.047)	(0.051)
IV	0.177	0.179	0.186	0.178	0.169
	(0.045)	(0.045)	(0.045)	(0.044)	(0.057)
RDD	0.056	0.058	0.043	0.038	0.027
	(0.053)	(0.053)	(0.051)	(0.051)	(0.061)
Observations	13,440	13,440	13,440	13,440	13,440
Pseudo R-squared	0.013	0.018	0.038	0.039	0.057
Top 5		Y	Y	Y	Y
Journal FE			Y	Y	Y
Reporting Method				Y	Y
Article Weights					Y

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% level. Robust standard errors are in parentheses, clustered by article. In columns 4 and 5, we control for the way author(s) report statistical significance, i.e., p-value, t-statistic or coefficient and standard error. In column 5, we use the inverse of the number of tests presented in the same article to weight observations.

Table 2: Caliper Test, Significant at the 5% Level

	(1) $Z > 1.96$	(2) $Z > 1.96$	(3) $Z > 1.96$	(4) $Z > 1.96$	(5) $Z > 1.96$
DID	0.051	0.051	0.054	0.058	0.096
	(0.033)	(0.033)	(0.037)	(0.036)	(0.042)
IV	0.071	0.069	0.066	0.073	0.093
	(0.031)	(0.031)	(0.035)	(0.035)	(0.043)
RDD	-0.005	-0.007	-0.004	0.001	0.018
	(0.035)	(0.048)	(0.036)	(0.069)	(0.039)
Observations	3,171	3,171	3,171	3,171	3,171
Pseudo R-squared	0.003	0.004	0.015	0.016	0.025
Top 5		Y	Y	Y	Y
Journal FE			Y	Y	Y
Reporting Method				Y	Y
Article Weights					Y
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% level. Robust standard errors are in parentheses, clustered by article. In columns 4 and 5, we control for the way author(s) report statistical significance, i.e., p-value, t-statistic or coefficient and standard error. In column 5, we use the inverse of the number of tests presented in the same article to weight observations.

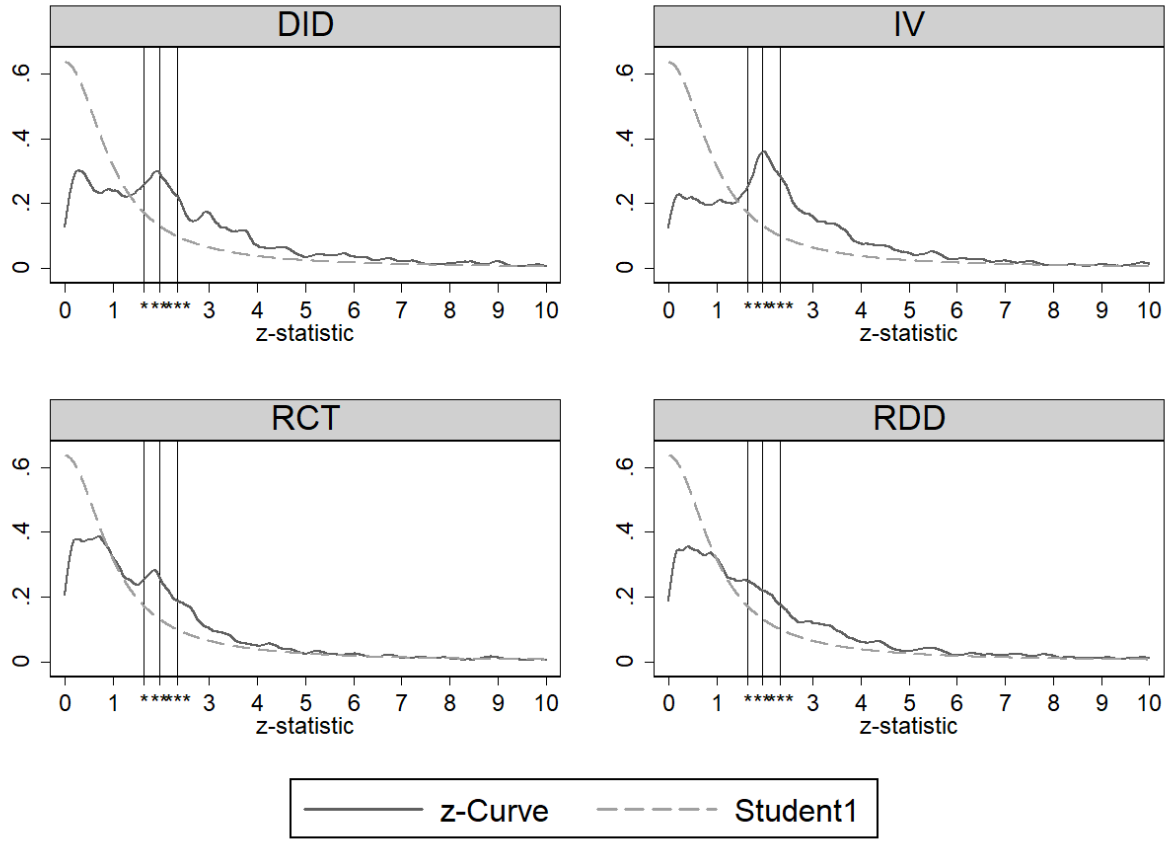
Table 3: Excess Test Statistics

z-Statistic	p-Value	DID	IV	RCT	RDD
(1.00-1.65)	(1.00-0.10)	-0.181	-0.238	-0.126	-0.098
(1.65-1.96)	(0.10-0.05)	0.038	0.038	0.035	0.019
(1.96-2.33)	(0.05-0.01)	0.070	0.101	0.044	0.037
(2.33-3.00)	(0.01-0.00)	0.073	0.098	0.048	0.041

This table displays the percentage of misallocated test statistics in each confidence interval. This table uses a Student(1) distribution and unweighted observed distributions. For example, 23.8% of IV estimates (for $z \in [1, 3]$) should be statistically insignificant. Of these missing tests almost half of them (10.1%) can be found in $z \in [1.96, 2.33]$.

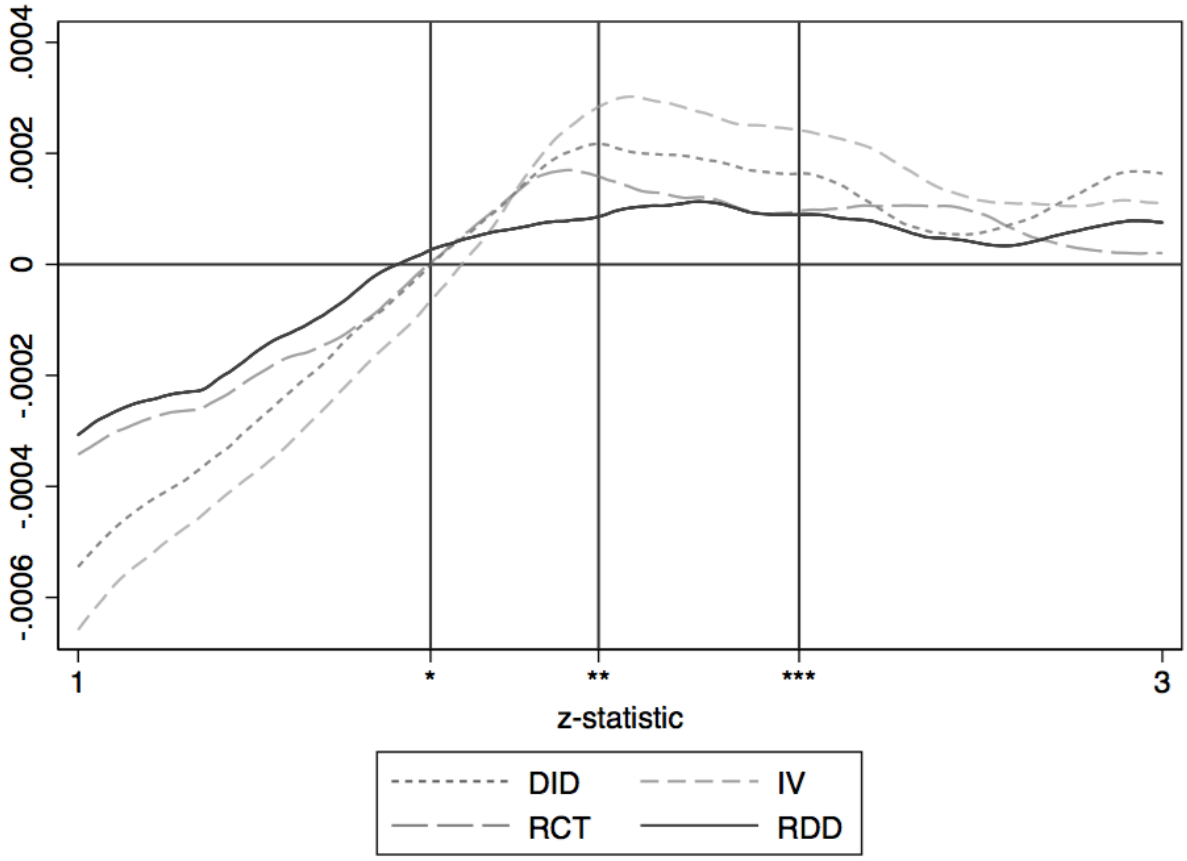
Appendices: NOT FOR PUBLICATION

Figure A1: Z-Curves and Student(1) Distribution



This figure displays the smoothed densities from Figures 1 and 2. The Student(1) distributions is used as a reference distribution to detect excess (or missing) tests. Reference lines are displayed at conventional significance levels.

Figure A2: Excess z-Statistics



This figure displays the "excess" test statistics for each of the four methods. Excess tests are the difference of the observed distribution (as in Figure 2) from a continuous monotonically decreasing distribution. For this figure we use the Student(1) distribution corresponding to Table 3 and (Brodeur et al., 2016). Reference lines are provided at conventional significance levels.

Table A1: Summary Statistics

Journal	DID	IV	RCT	RDD	Articles	Tests
AEJ: Applied Economics	5	5	9	3	22	1,256
AEJ: Economic Policy	9	5	2	5	21	991
AEJ: Macroeconomics	0	2	0	0	2	20
American Economic Review	6	11	8	2	27	1,114
Econometrica	1	2	0	1	4	185
Economic Journal	6	9	0	1	16	877
Economic Policy	0	1	0	0	1	6
Experimental Economics	0	2	1	0	3	29
Journal of Applied Econometrics	0	0	0	1	1	102
Journal of Development Economics	7	5	8	2	22	911
Journal of Economic Growth	0	3	0	0	3	23
Journal of Finance	3	8	3	2	16	887
Journal of Financial Economics	6	9	0	2	17	290
Journal of Financial Intermediation	5	5	0	1	11	283
Journal of Human Resources	0	6	2	3	11	630
Journal of International Economics	2	4	0	0	6	267
Journal of Labor Economics	3	3	4	1	11	429
Journal of Political Economy	0	2	2	1	5	451
Journal of Public Economics	11	6	6	6	29	1,225
Journal of the European Economic Association	4	2	3	0	9	292
Journal of Urban Economics	6	4	0	1	11	610
Review of Economic Studies	0	3	1	0	4	200
Review of Financial Studies	9	4	0	3	16	646
The Quarterly Journal of Economics	1	4	3	3	11	571
The Review of Economics and Statistics	7	9	7	6	29	1,145
Total Articles	91	114	59	44	308	
Total Tests	3,089	2,749	3,536	4,066		13,440

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 A2: Excess Test Statistics - Cauchy Distribution

z-Statistic	p-Value	DID	IV	RCT	RDD
(1.00-1.65)	(1.00-0.10)	-0.230	-0.286	-0.175	-0.147
(1.65-1.96)	(0.10-0.05)	0.047	0.047	0.043	0.028
(1.96-2.33)	(0.05-0.01)	0.084	0.116	0.058	0.052
(2.33-3.00)	(0.01-0.00)	0.098	0.124	0.073	0.067

This table displays the percentage of misallocated tests in each confidence interval. This table uses a Cauchy(0.5) distribution and unweighted observed distributions. For example, 28.6% of tests (for $z \in [1, 3]$) for IV should be statistically insignificant. Of these missing tests, more than a third of them (11.6%) can be found in $z \in [1.96, 2.33]$.

Table A3: Significant at the 10% Level

	(1) $Z > 1.65$	(2) $Z > 1.65$	(3) $Z > 1.65$	(4) $Z > 1.65$	(5) $Z > 1.65$
DID	0.126	0.127	0.127	0.122	0.149
	(0.038)	(0.038)	(0.041)	(0.041)	(0.045)
IV	0.185	0.186	0.191	0.181	0.167
	(0.040)	(0.040)	(0.040)	(0.040)	(0.051)
RDD	0.038	0.040	0.024	0.017	0.018
	(0.048)	(0.048)	(0.047)	(0.047)	(0.055)
Observations	13,440	13,440	13,440	13,440	13,440
Pseudo R-squared	0.015	0.016	0.037	0.040	0.049
Top 5		Y	Y	Y	Y
Journal FE			Y	Y	Y
Reporting Method				Y	Y
Article Weights					Y

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% level. Robust standard errors are in parentheses, clustered by article. In columns 4 and 5, we control for the way author(s) report statistical significance, i.e., p-value, t-statistic or coefficient and standard error. In column 5, we use the inverse of the number of tests presented in the same article to weight observations.

Table A4: Significant at the 1% Level

	(1) $Z > 2.33$	(2) $Z > 2.33$	(3) $Z > 2.33$	(4) $Z > 2.33$	(5) $Z > 2.33$
DID	0.120	0.124	0.121	0.123	0.154
	(0.048)	(0.047)	(0.047)	(0.050)	(0.055)
IV	0.150	0.154	0.157	0.156	0.139
	(0.049)	(0.048)	(0.045)	(0.047)	(0.060)
RDD	0.068	0.073	0.058	0.053	0.024
	(0.055)	(0.055)	(0.050)	(0.053)	(0.062)
Observations	13,440	13,440	13,440	13,434	13,434
Pseudo R-squared	0.009	0.010	0.039	0.040	0.065
Top 5		Y	Y	Y	Y
Journal FE			Y	Y	Y
Reporting Method				Y	Y
Article Weights					Y

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% level. Robust standard errors are in parentheses, clustered by article. In columns 4 and 5, we control for the way author(s) report statistical significance, i.e., p-value, t-statistic or coefficient and standard error. In column 5, we use the inverse of the number of tests presented in the same article to weight observations.

Table A5: Covariates

	(1) $Z > 2.33$	(2) $Z > 1.96$	(3) $Z > 1.65$
Top 5	-0.047	-0.085	-0.082
	(0.129)	(0.114)	(0.103)
Finance	-0.059	-0.075	-0.089
	(0.128)	(0.112)	(0.100)
Macroeconomics	-0.034	0.023	0.090
	(0.148)	(0.109)	(0.097)
General Interest	-0.057	-0.083	-0.089
	(0.121)	(0.107)	(0.095)
Experimental	-0.130	-0.184	-0.146
	(0.126)	(0.119)	(0.119)
Development	-0.137	-0.180	-0.155
	(0.151)	(0.136)	(0.120)
Labor	-0.020	-0.037	-0.017
	(0.076)	(0.083)	(0.083)
Public	-0.006	-0.035	-0.035
	(0.130)	(0.118)	(0.106)
Urban	-0.338	-0.369	-0.335
	(0.125)	(0.110)	(0.098)
T-Statistics	0.093	0.097	0.122
	(0.097)	(0.098)	(0.093)
P-Value	-0.053	-0.006	0.035
	(0.054)	(0.067)	(0.070)
Constant	0.358	0.472	0.542
	(0.126)	(0.113)	(0.101)
Observations	13,451	13,451	13,451
R-squared	0.045	0.047	0.040
Top 5	Y	Y	Y
Article Weights	Y	Y	Y

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% level. In column 2, the dependent variable is a dummy for whether the test statistic is significant at the 5% level. In column 3, the dependent variable is a dummy for whether the test statistic is significant at the 10% level. Robust standard errors are in parentheses, clustered by article.

Table A6: Caliper Test, Significant at the 10% Level

	(1) $Z > 1.65$	(2) $Z > 1.65$	(3) $Z > 1.65$	(4) $Z > 1.65$	(5) $Z > 1.65$
DID	0.044	0.044	0.053	0.051	0.066
	(0.033)	(0.033)	(0.034)	(0.035)	(0.037)
IV	0.091	0.090	0.090	0.087	0.084
	(0.035)	(0.035)	(0.037)	(0.037)	(0.042)
RDD	-0.056	-0.057	-0.053	-0.056	-0.002
	(0.028)	(0.029)	(0.032)	(0.032)	(0.035)
Observations	3,239	3,239	3,239	3,239	3,239
Pseudo R-squared	0.009	0.009	0.016	0.016	0.015
Top 5		Y	Y	Y	Y
Journal FE			Y	Y	Y
Reporting Method				Y	Y
Article Weights					Y
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% level. Robust standard errors are in parentheses, clustered by article. In columns 4 and 5, we control for the way author(s) report statistical significance, i.e., p-value, t-statistic or coefficient and standard error. In column 5, we use the inverse of the number of tests presented in the same article to weight observations.

Table A7: Caliper Test, Significant at the 1% Level

	(1) $Z > 2.33$	(2) $Z > 2.33$	(3) $Z > 2.33$	(4) $Z > 2.33$	(5) $Z > 2.33$
DID	-0.024	-0.022	-0.026	-0.028	-0.004
	(0.031)	(0.031)	(0.034)	(0.034)	(0.032)
IV	-0.013	-0.009	-0.004	-0.008	-0.022
	(0.033)	(0.033)	(0.036)	(0.036)	(0.040)
RDD	-0.004	0.001	0.003	-0.001	-0.025
	(0.032)	(0.031)	(0.036)	(0.036)	(0.040)
Observations	2,722	2,722	2,722	2,722	2,722
Pseudo R-squared	0.000	0.001	0.007	0.008	0.015
Top 5		Y	Y	Y	Y
Journal FE			Y	Y	Y
Reporting Method				Y	Y
Article Weights					Y
Window Width	$Z > 1.83$ & $Z < 2.83$				

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% level. Robust standard errors are in parentheses, clustered by article. In columns 4 and 5, we control for the way author(s) report statistical significance, i.e., p-value, t-statistic or coefficient and standard error. In column 5, we use the inverse of the number of tests presented in the same article to weight observations.