Online Appendix to: Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics

By Abel Brodeur, Nikolai Cook, and Anthony Heyes

This file contains the online appendices noted in Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics

A. For Online Publication

Randomization Tests

In Appendix Figures A23, A24 and A25, we check the robustness of our randomization tests - in particular to the likely high correlation between specifications in any given table. For a review, these tests have a null hypothesis that there is an equal proportion of tests just above and below some threshold. We use a window of half-width 0.25. This means that for the first significance threshold (pictured in Appendix Figure A23) we use only test statistics between $1.4 < z < 1.9$ (1.65 ± 0.25). From this set, we randomly select only one test from each table. By method, we then conduct the randomization test 100 times. The figures present histograms of the p-value. In Appendix Figure A23, we show that 84 of the 100 'bootstrap' samples had a p-value less than 0.05. Specifically, we could reject the null hypothesis that the proportion of significant test statistics is equal to or lower than the proportion of insignificant test statistics (one way test) for the conventional 10% threshold. The average number of DID test statistics used is 254. In all cases, IV had significantly more test statistics above the 10% threshold than below, even with the reduced sample size. RCT and RDD fared slightly better.

The main result is presented in Appendix Figure A24. Here, we test whether the proportion of test statistics in $[1.96, 2.21]$ is greater than the proportion of test statistics in $[1.71, 1.96)$. In none of the 100 replications for DID, RCT and RDD did the randomization test return a statistically significant result. For 40% of IV samples, there was a statistically significant larger proportion of tests above rather than below $z = 1.96$, despite the reduced sample size to an average of 324 tests.

Another interesting result is found in Appendix Figure A25. Again, we test whether the proportion of tests above a threshold ($z = 2.58$) is greater than below. For no method, and in no sample, was there a statistically significant result. At this high level of statistical significance, there does not seem to be any disproportional amount of test statistics above the threshold.

Probit Analysis: Full Sample

In this subsection, we use probit regressions to study whether the likelihood that a test delivers a significant result is related to the method employed. The main difference between this analysis and our caliper test analysis is that we do not restrict the sample to a narrow band around arbitrary statistical significance thresholds. In other words, the following probit estimates are estimated on the full range of p-values. This approach thus has little to say about the extent of p-hacking, but remains informative as published studies relying on RCT or RDD may be more likely to report tightly-estimated zeros.

Appendix Table A29 presents estimates of Equation 2 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).

In the most parsimonious specification, we find that DID and IV estimates are about 16 and 21% more likely to be statistically significant than a RCT estimate, respectively. Our estimates are statistically significant at the 1 percent level. RDD estimates are also positive, but smaller in magnitude.

In column 2, we enrich our specifications with authors and articles' characteristics. In column 3, we include field fixed effects. Column 4 includes journal fixed effects. Our conclusions remain unchanged. For instance, our IV estimates remain statistically significant at the 1 percent level across specifications and range from 18% to 19%.

Appendix Tables A30 and A31 replicate Appendix Table A29 for the other conventional significance levels. For the 1.65 cut-off, the estimates and conclusions are similar to the 5% significance threshold, while the estimates are all smaller for the 1% cut-off.

Robustness Checks: Caliper Test

In Appendix Table A22 we showed that the level of p-hacking is roughly the same if we restricted the analysis to only the first results table in each article. However, it is not fully clear that restricting to the first table solves the issue that the number of tables (and test statistics) between methods is very different. Another interesting robustness check is to randomly sample (50% of) the t statistics within each paper and conduct the same caliper tests as in the full sample. In Appendix Figure A26 we present the results of bootstrapping the caliper tests by randomly sampling t statistics within papers 1000 times.[1] These would coincide with the coefficients from Appendix Table A22, in the first column (no controls). The left panel displays histograms of the p-values for each of the regression coefficients. The right panel displays histograms of the associated estimated effect sizes. We find DID and IV reject the null hypothesis that the probability of

[1] In Appendix Figure A27 we present the results of bootstrapping our sample by randomly including papers - a different re-sampling margin - results are similar.

being statistically significant is greater than RCT (the omitted method in the regression). The estimated coefficients also vary around means roughly centered to those in Appendix Table A22.

In Appendix Table A32, we check whether our findings are robust to coding/data collection methods. As already mentioned, we replicated the work of each other and ended up collecting the same tests for the vast majority of research articles. We drop the articles for which we could not easily reach an agreement on which tests to select. The point estimates are very similar to our baseline specification.

In Appendix Table A33, we test whether the main findings are not driven by journal articles relying on multiple methods.[2] More than 10 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. We find that IV (and to some extent DID) articles report more marginally significant tests at the 5 percent level than RCT articles.

Thus far, we have relied on all journals in our sample. As an additional robustness check, we explore the sensitivity of our results to the omission of a subset of journals. In Appendix Table A35 , 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 tests. Most of our estimates are statistically significant at conventional levels for IV and range from 5 to 8 percentage points when the full set of controls and journal fixed effects are included. This suggests that our findings are unlikely to be driven by practice in specific fields or journals.

P-Hacking and Instrumental Variables

Interestingly, we find that the degree of p-hacking in the second stage is related to the strength of the first stage. The median F-statistic in our sample is just over 30; we divide them into below and above the median. See Figure 5b. The left panel displays the distribution of tests for IVs with a relatively low F-statistic, while the right panel displays the distribution of tests for IVs with a relatively high F-statistic. Because not all IV estimates have an associated F-statistic, a total of 1,414 statistics are used in this analysis. The left panel contains 681 tests, while the right contains 733 tests. Both panels continue to display a large increase

[2]Similarly, our main findings are robust to restricting the sample to authors with multiple papers. See Appendix Figure A28. Appendix Table A34 reports summary statistics for these authors.

in mass just at the statistical threshold levels. The remarkable difference is that second stage results from relatively 'weak' IVs have a much higher proportion centered around the conventional significance thresholds.

When researchers are skeptical of instrumental variables, it is often because they are unconvinced that the exclusion restriction holds. We can address this concern by comparing treatment effect on the treated results from RCT studies, i.e., RCT papers with partial compliance, to IV in observational studies. This is an interesting exercise since the exclusion restriction is more likely to hold when the instrument is randomly assigned. In Appendix Figure A17, the left panel displays the distribution of test statistics for IV tests in RCT studies. The right panel displays the distribution of tests for IV results from observational studies. We see that IV results from RCT studies display a markedly smaller degree of p-hacking than IV from observational studies. This points us to suspect that the application or reception of IV's - rather than the methodology itself - is generating this curve. The total number of tests graphed is 6,110 (since we use a cut-off of $t < 10$). The left panel displays 1,154 observations (almost 20% of the sample) while the right displays 4,956 observations.

We also present a related exercise in which we compare IV test statistics in RCT papers to RCT test statistics. Admittedly this is an unbalanced sample - many RCT studies do not report IV estimates to cope with partial compliance. Appendix Figure A18 displays histograms of test statistics in published papers for $z \in [0, 10]$. The left panel displays tests of instrumental variables in RCT studies. The right panel displays tests for IV in observational studies, i.e., non-RCT studies. This figure shows that the distribution of test statistics is quite similar in these two subsamples.

Appendix Figures and Tables

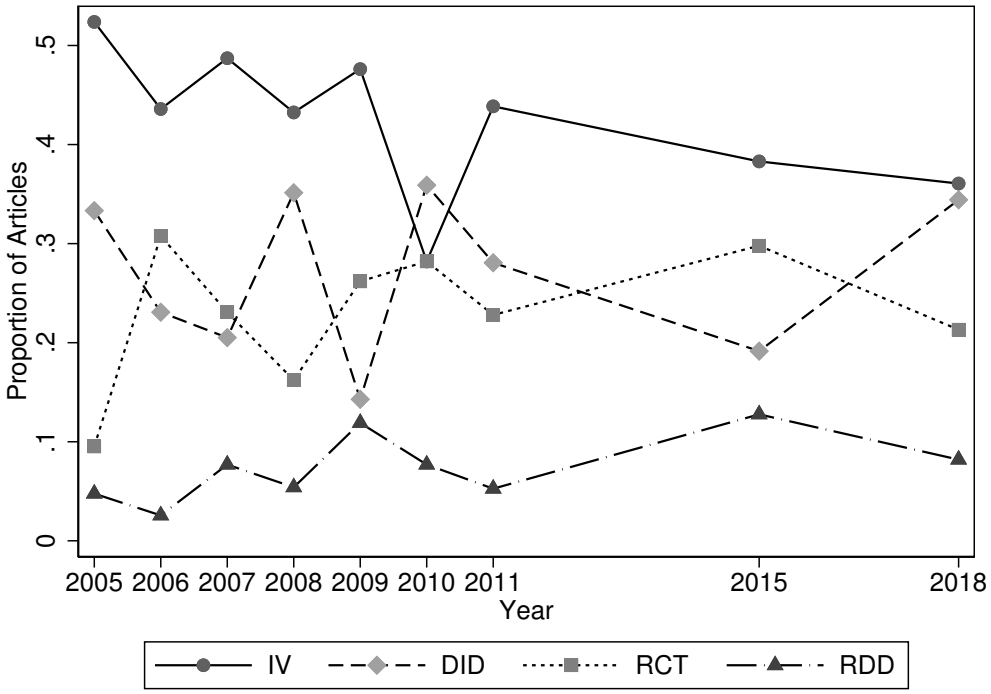

Figure A1. Proportion of Articles by Method Over Time for Three Journals

Note: This figure illustrates the proportion of articles by method over the time period 2005–2011, 2015 and 2018 for the *American Economic Review*, *Journal of Political Economy* and the *Quarterly Journal of Economics*.

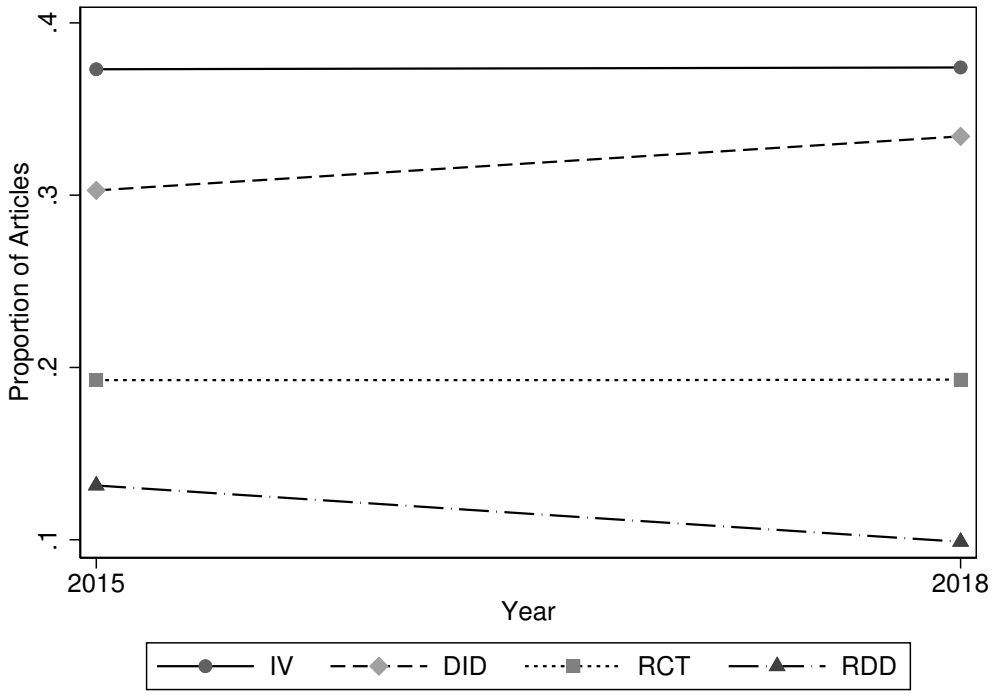

Figure A2. Proportion of Articles by Method Over Time (Top 25)

Note: This figure illustrates the proportion of articles by method over the time period 2015 and 2018 for the top 25 journals in economics and finance.

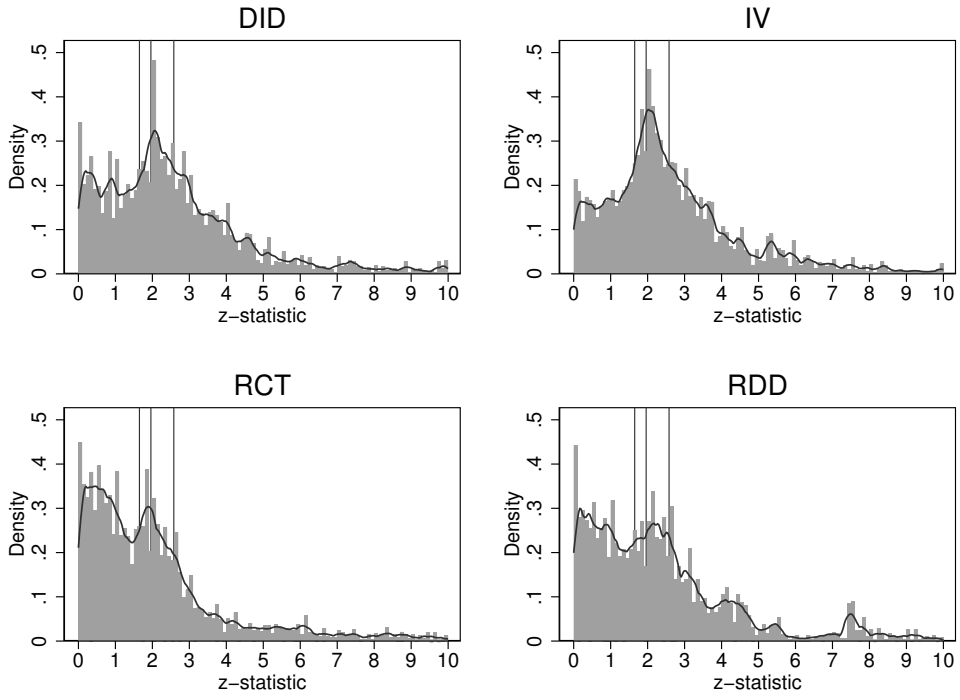

FIGURE A3. z-STATISTICS BY METHOD: WEIGHTED

Note: 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.

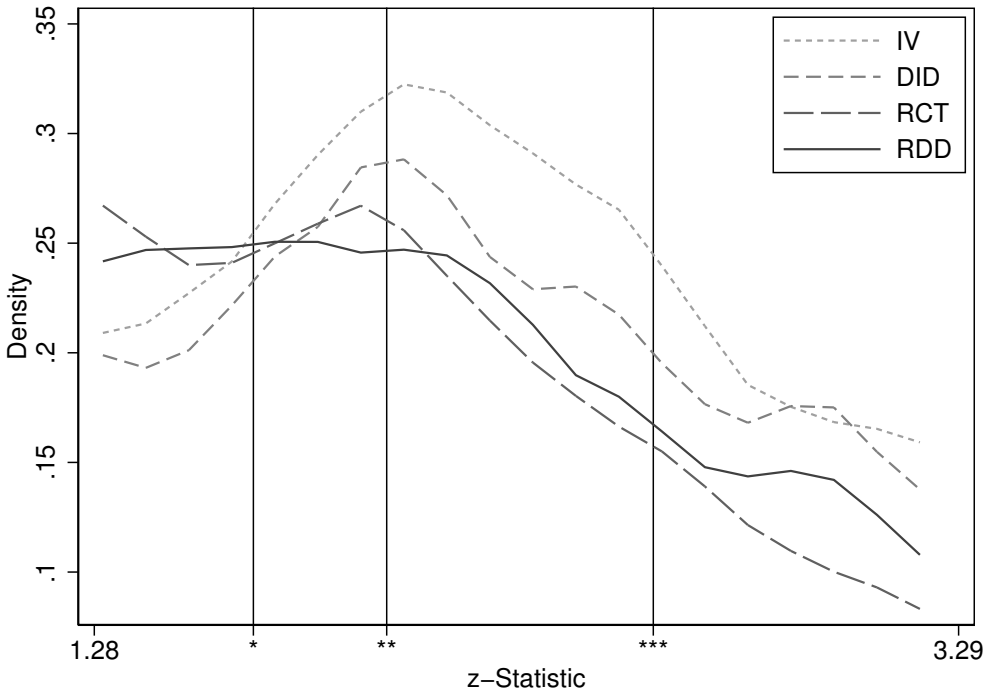

FIGURE A4. z-STATISTICS BY METHOD: SMOOTHED DENSITIES

Note: This figure displays the smoothed densities (Epanechnikov) from Figure 2 for $1.28 < z < 3.29$. A density is displayed for each of four methods. Reference lines are displayed at the conventional two-tailed significance levels. No weights applied.

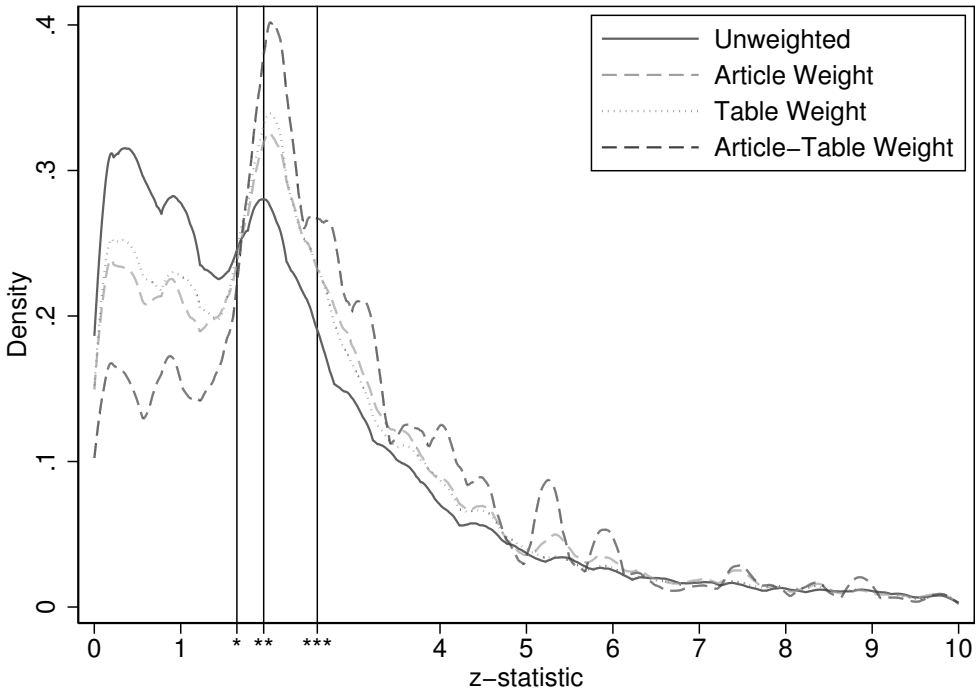

FIGURE A5. z-STATISTICS BY WEIGHTING SCHEME

Note: 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.

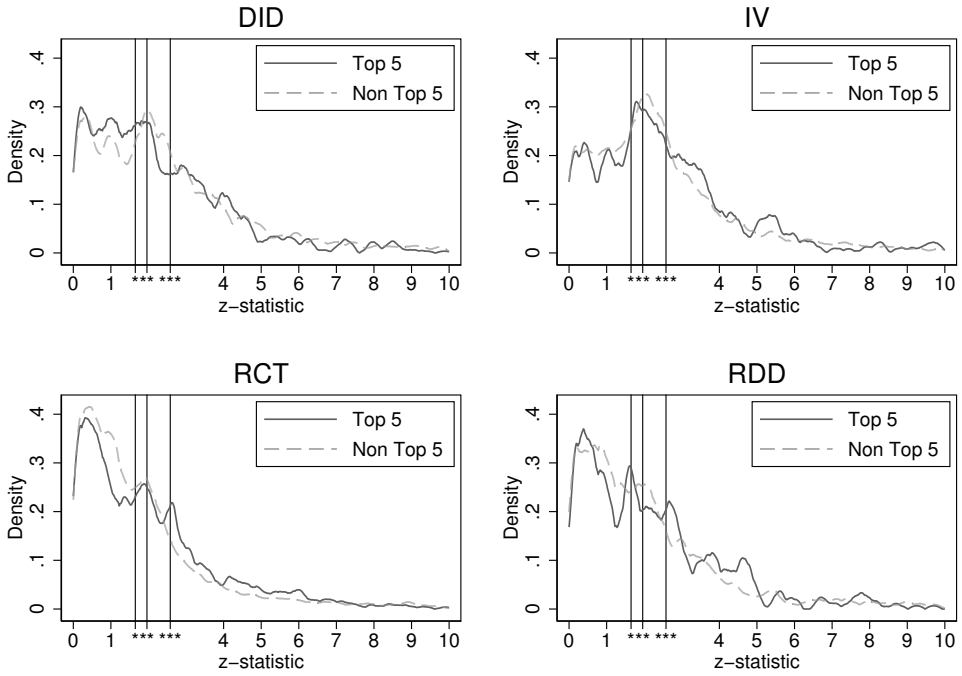

Figure A6. z-Statistics by Method and Journal Ranking

Note: 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 published in the top 5. Lines in light gray (dashes) are for articles published in non-top 5. 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. No weights applied.

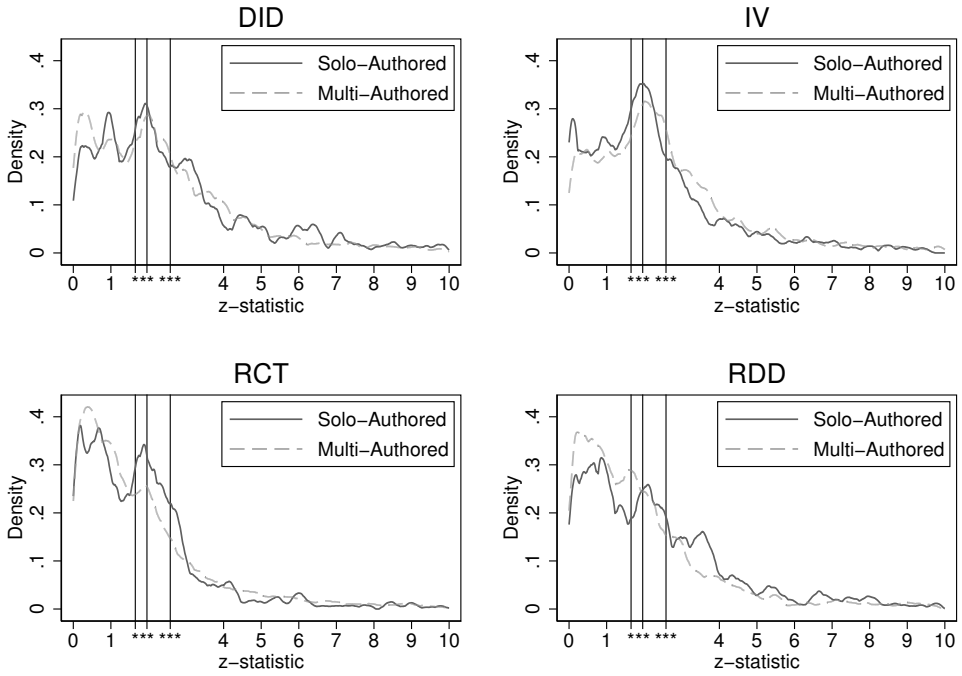

FIGURE A7. z-STATISTICS BY METHOD AND NUMBER OF AUTHORS

Note: 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. No weights have been applied.

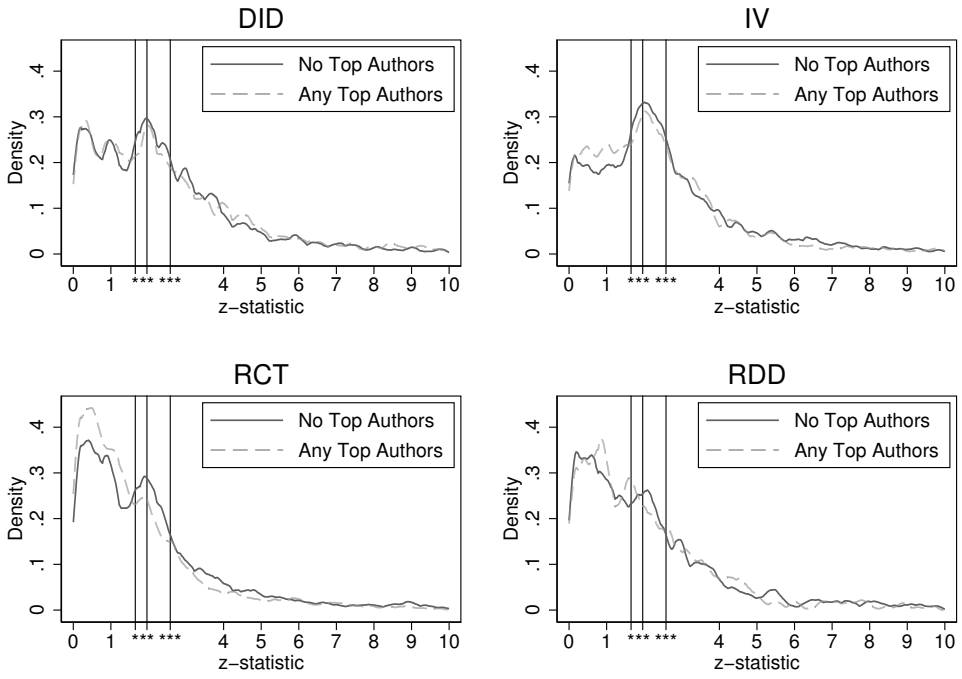

FIGURE A8. z-STATISTICS BY METHOD AND AFFILIATION

Note: 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. No weights have been applied.

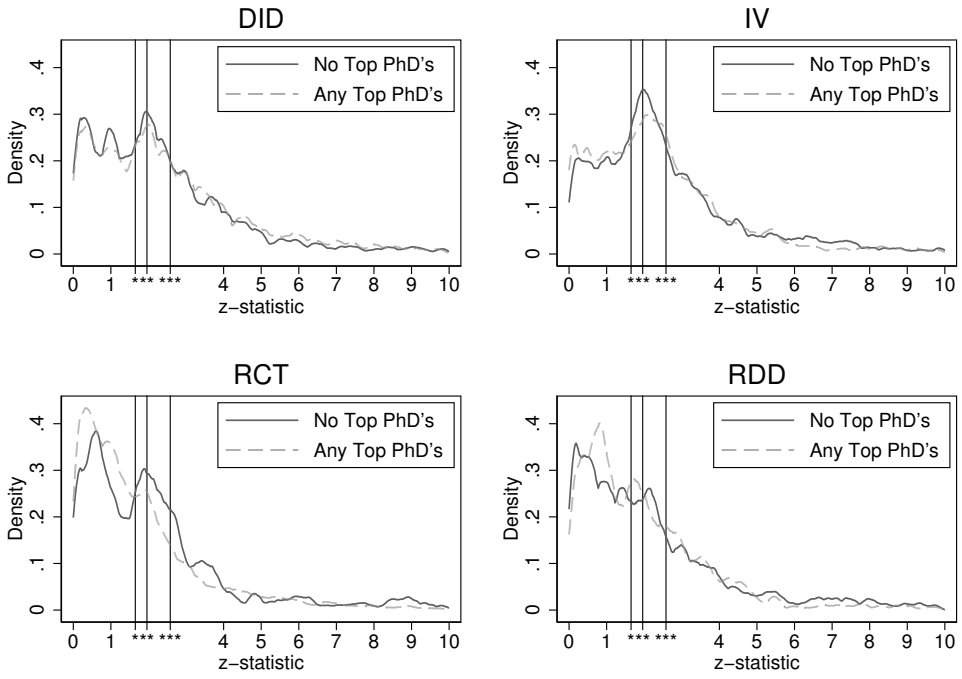

FIGURE A9. z-STATISTICS BY METHOD AND PHD INSTITUTION

Note: 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. No weights have been applied.

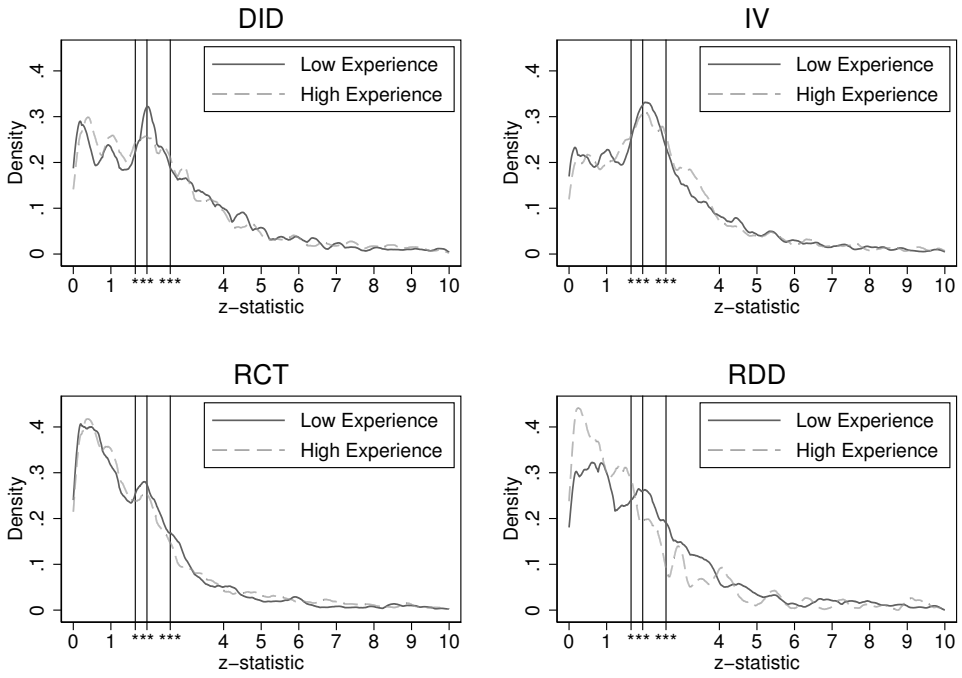

FIGURE A10. z-STATISTICS BY METHOD AND YEARS OF EXPERIENCE

Note: 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. No weights have been applied.

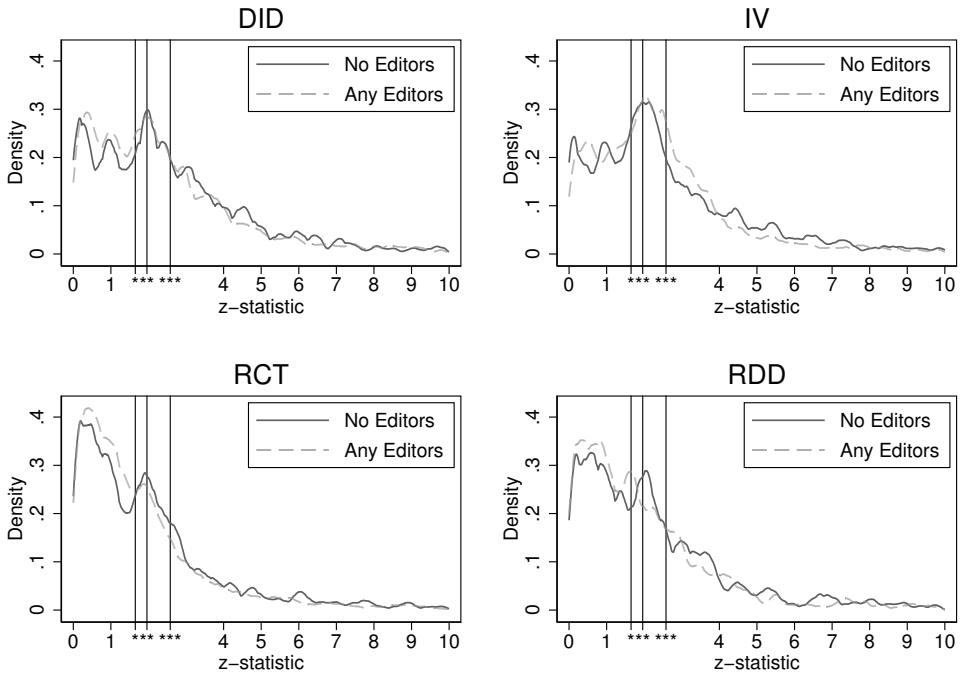

FIGURE A11. z-STATISTICS BY METHOD AND EDITOR

Note: 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. No weights have been applied.

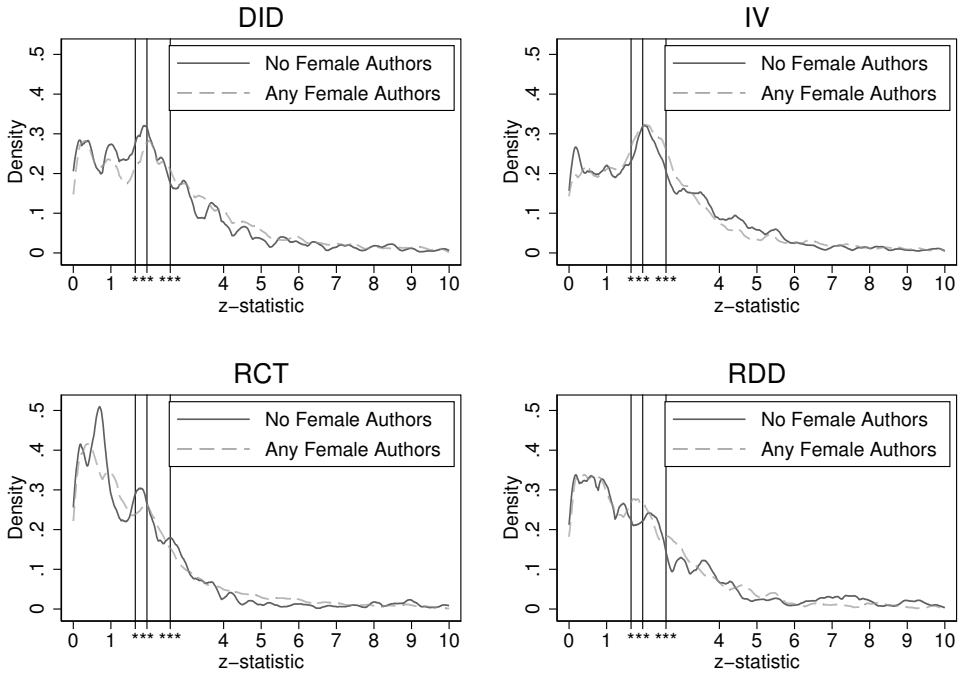

FIGURE A12. z-STATISTICS BY METHOD AND AUTHORS' GENDER

Note: 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. No weights have been applied.

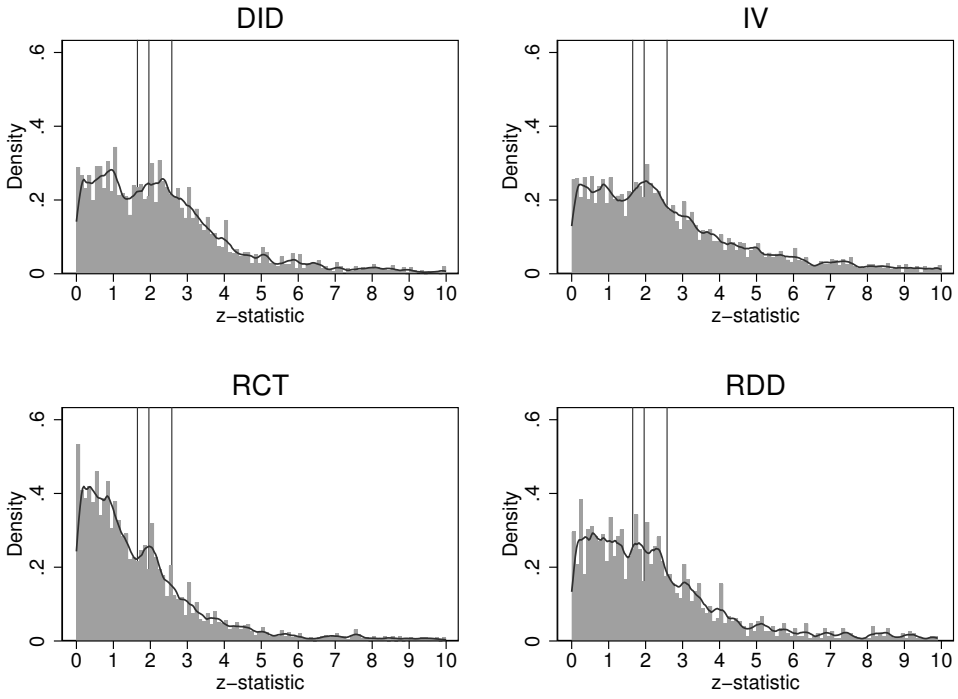

FIGURE A13. BRODEUR ET AL. (2016) SAMPLE BY METHOD

Note: This figure displays histograms of test statistics for $z \in [0, 10]$ for the *American Economic Review*, *Journal of Political Economy* and the *Quarterly Journal of Economics* from 2005–2011. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. No weights have been applied.

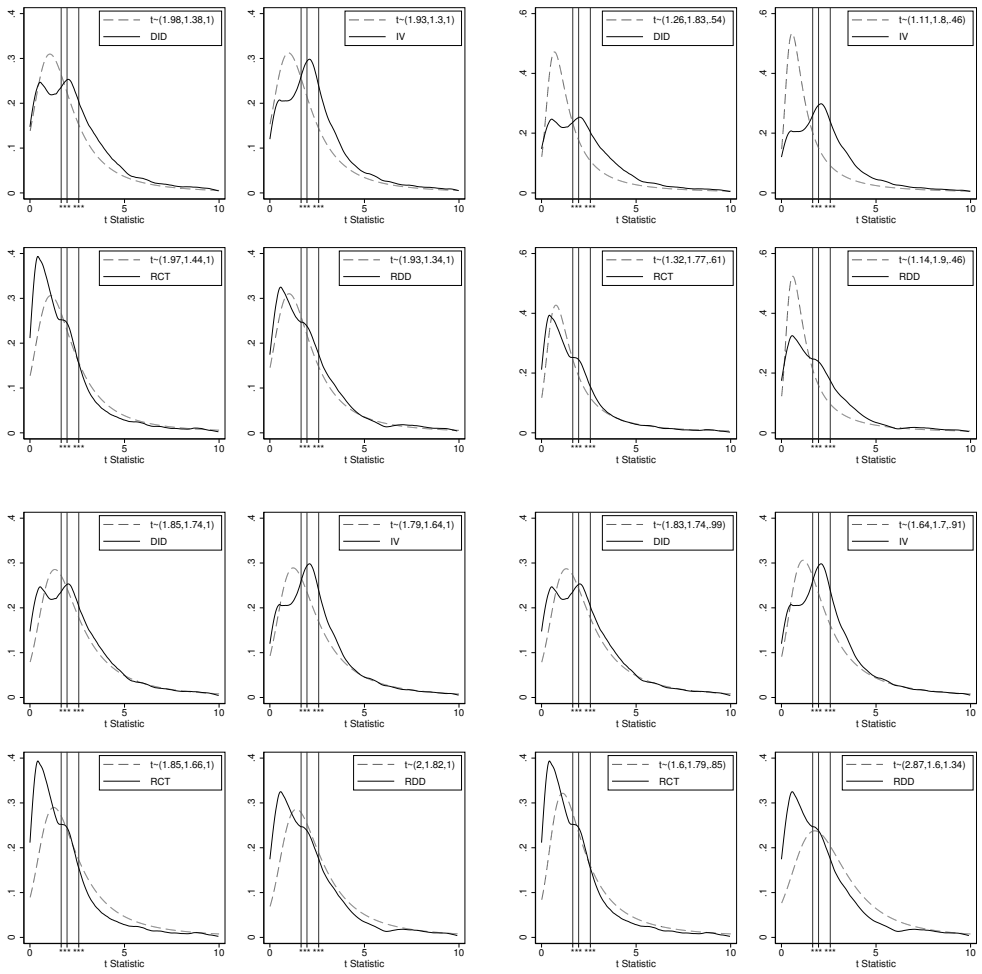

Figure A14. Excess Test Statistics by Method Using Maximum Likelihood, Alternative Thresholds and a Generalized t Distribution

Note: This figure presents an alternative approach to that presented in Figure 4. This figure presents t distributions fit to the observed distribution's tails by maximum likelihood. Panel A and B use a threshold of $z = 3$. Panel C and D use a threshold of $z = 5$. Panel A and C use a non-central t distribution (degrees of freedom and non-centrality parameter). Panel B and D use a generalized t distribution which includes a third parameter for scale. No weights have been applied.

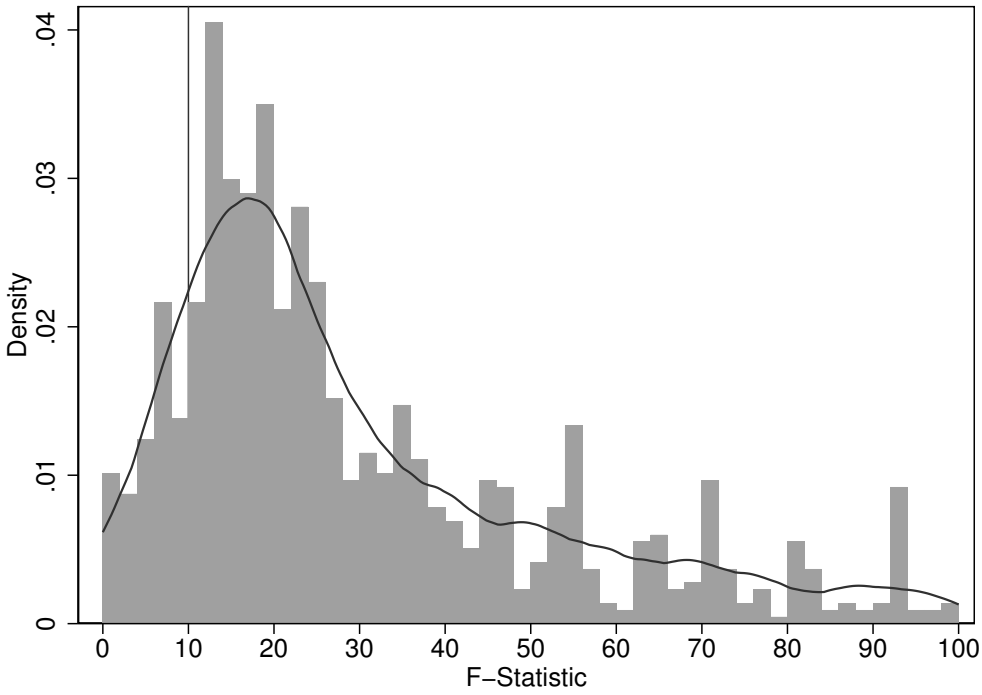

FIGURE A15. INSTRUMENTAL VARIABLE: FIRST STAGE F-STATISTICS

Note: This figure displays an histogram of First Stage F-Statistics of instrumental variables for $F \in [0, 100]$. No weights have been applied.

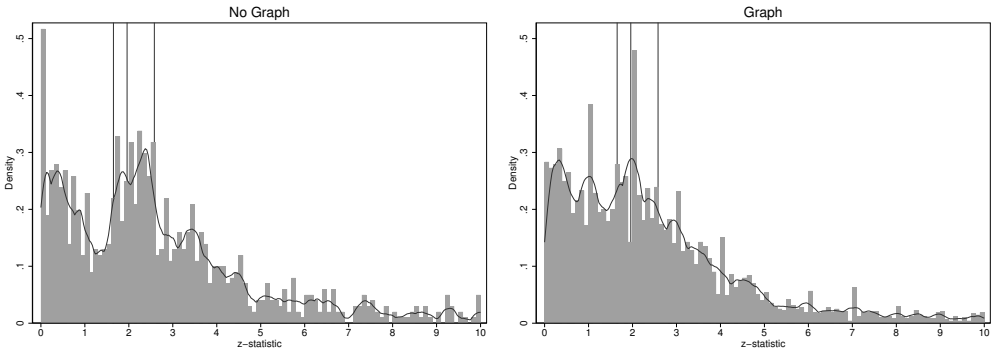

FIGURE A16. ROLE OF EVENT-STUDY GRAPHS FOR DID

Note: his figure displays histograms of test statistics for $z \in [0, 10]$. Panel A restricts the sample to DID articles without an event-study graph. Panel B restricts the sample to DID articles with an event-study graph. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. No weights have been applied.

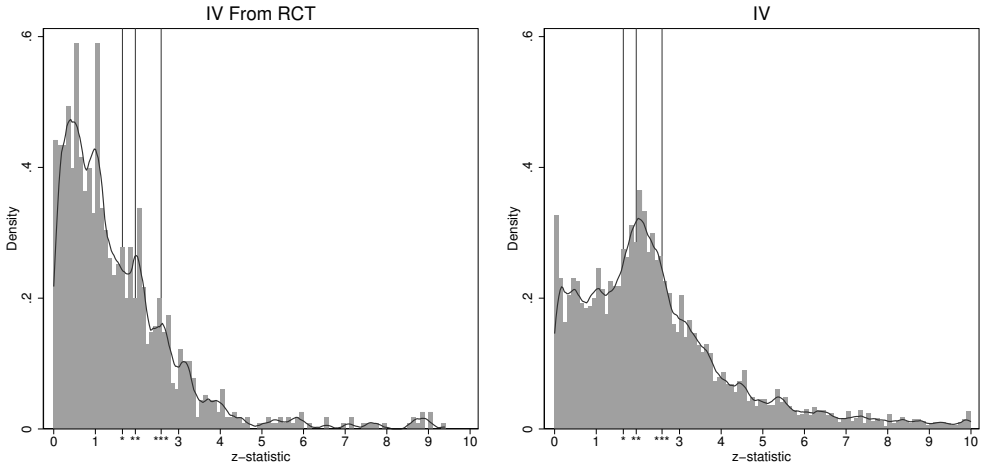

FIGURE A17. HISTOGRAM FOR IV TEST STATISTICS IN RCT AND NON-RCT STUDIES

Note: This figure displays histograms of test statistics in published papers for $z \in [0, 10]$. The left panel displays tests of instrumental variables in RCT studies. The right panel displays tests for IV in observational studies, i.e., non-RCT studies. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. No weights have been applied.

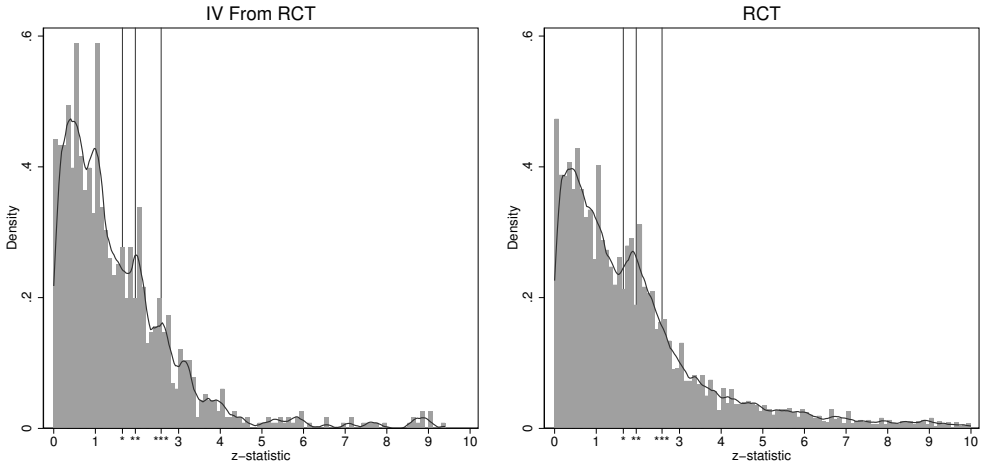

FIGURE A18. HISTOGRAM FOR RCT AND IV TEST STATISTICS IN RCT STUDIES

Note: This figure displays histograms of test statistics in published papers for $z \in [0, 10]$. The left panel displays tests of instrumental variables in RCT studies. The right panel displays tests for RCT. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. No weights have been applied.

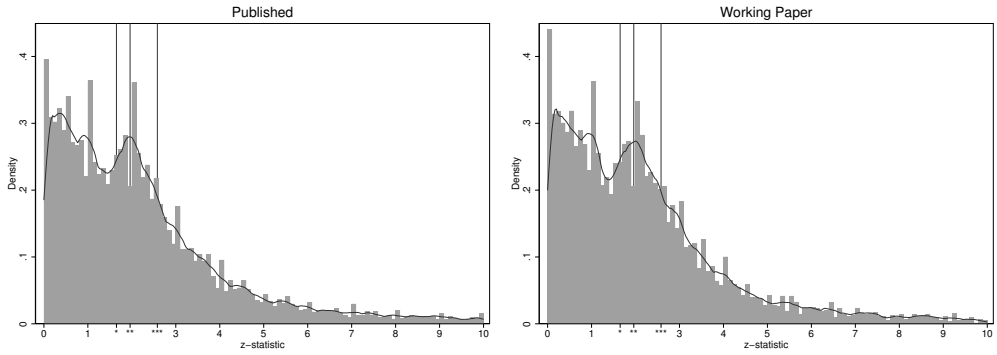

FIGURE A19. HISTOGRAM BY PUBLICATION STATUS - UNBALANCED SAMPLE

Note: This figure displays histograms of test statistics for $z \in [0, 10]$. Panel A restricts the sample to journal articles. Panel B restricts the sample to working papers. The sample includes all published articles in our sample, i.e., includes papers for which we did not find a working paper. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. No weights have been applied.

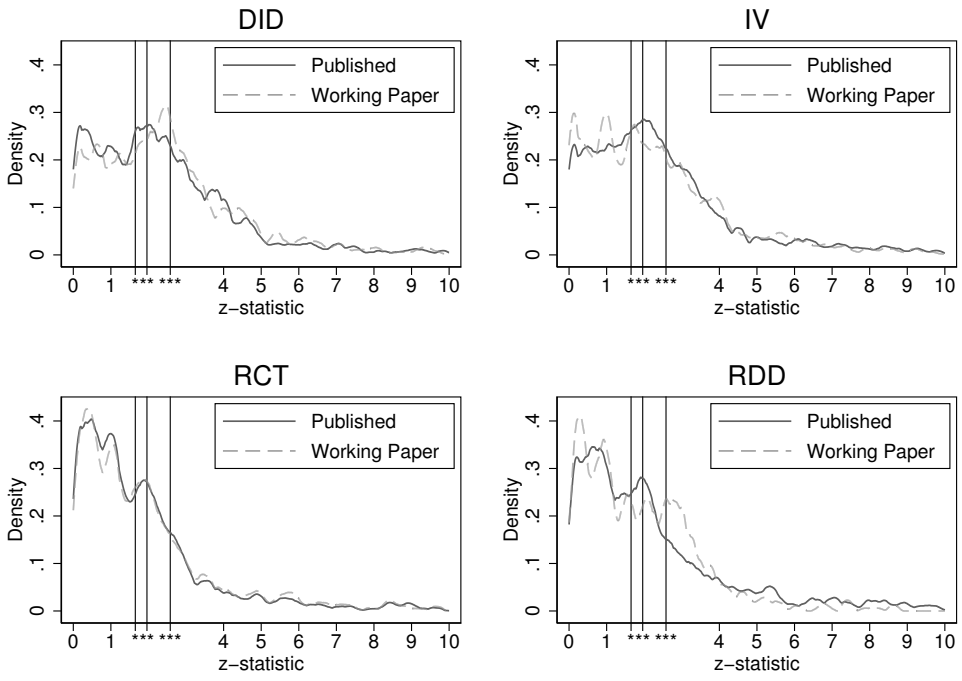

Figure A20. Histogram by Publication Status and Method - Balanced Sample

Note: This figure displays histograms of test statistics for $z \in [0, 10]$. Panel A restricts the sample to journal articles. Panel B restricts the sample to working papers. The sample is accordingly restricted to estimates from published articles that had an associated working paper. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. No weights have been applied.

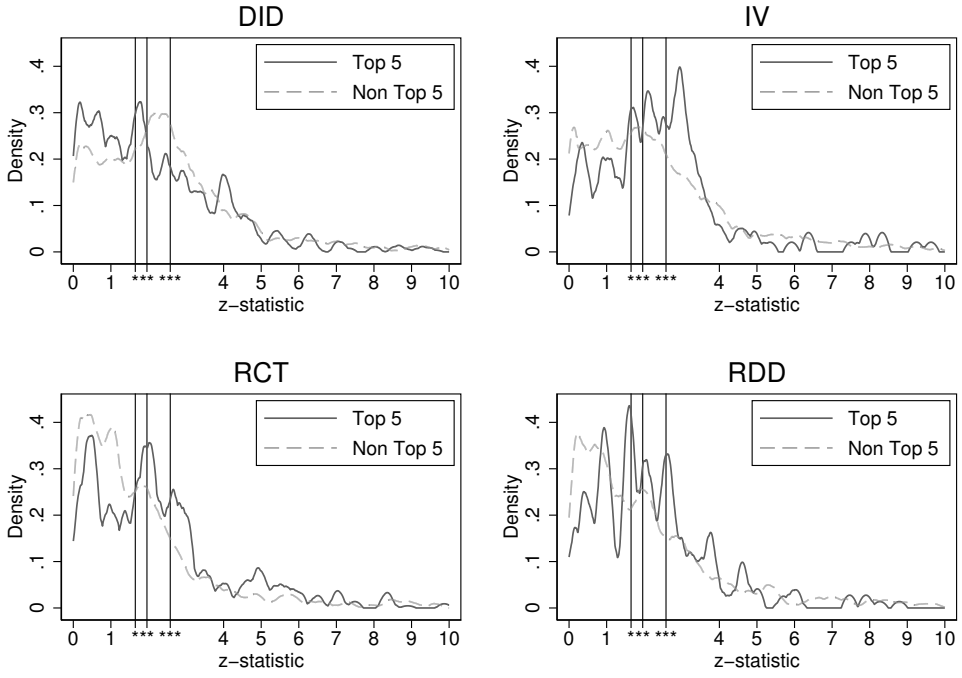

FIGURE A21. HISTOGRAM BY METHOD FOR TOP 5 AND NON-TOP 5 - WORKING PAPERS

Note: This figure displays histograms of test statistics in working papers for $z \in [0, 10]$ by method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT), and regression discontinuity design (RDD). Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. No weights have been applied.

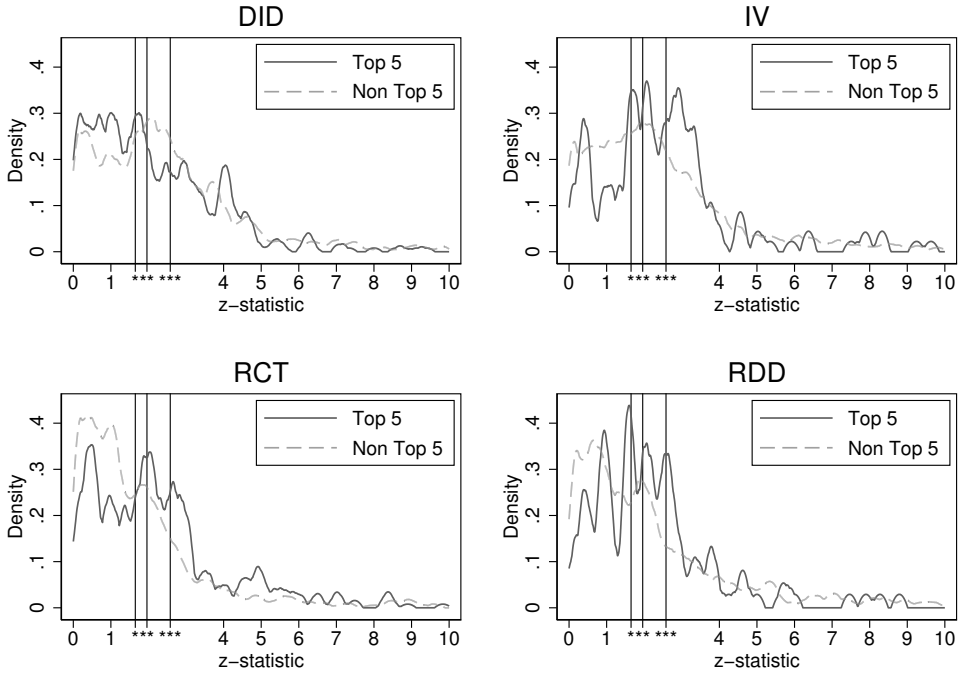

Figure A22. Histogram by Method for Top 5 and non-Top 5 - Published Papers (Balanced Sample)

Note: This figure displays histograms of test statistics in published papers for $z \in [0, 10]$ by method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT), and regression discontinuity design (RDD). The sample includes only published articles for which we did find a working paper. Lines in dark gray are for articles published in the top 5. Lines in light gray (dashes) are for articles published in non-top 5. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. No weights have been applied.

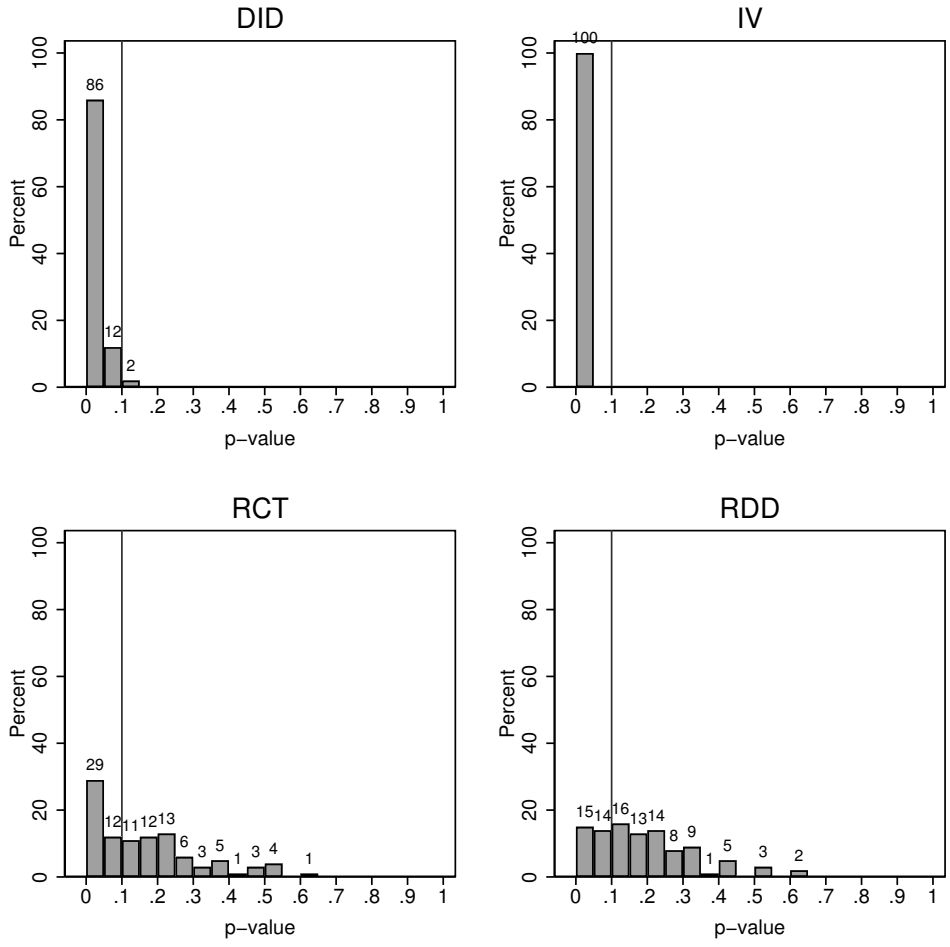

FIGURE A23. BOOTSTRAP RANDOMIZATION TESTS $z = 1.65$

Note: Tests have a null hypothesis that there is an equal proportion of tests just above and below $z = 1.65$. We use a window of half-width 0.25. From this set, we randomly select only one test from each table. By method, we then conduct the randomization test 100 times. The figures present histograms of the p-value. No weights have been applied.

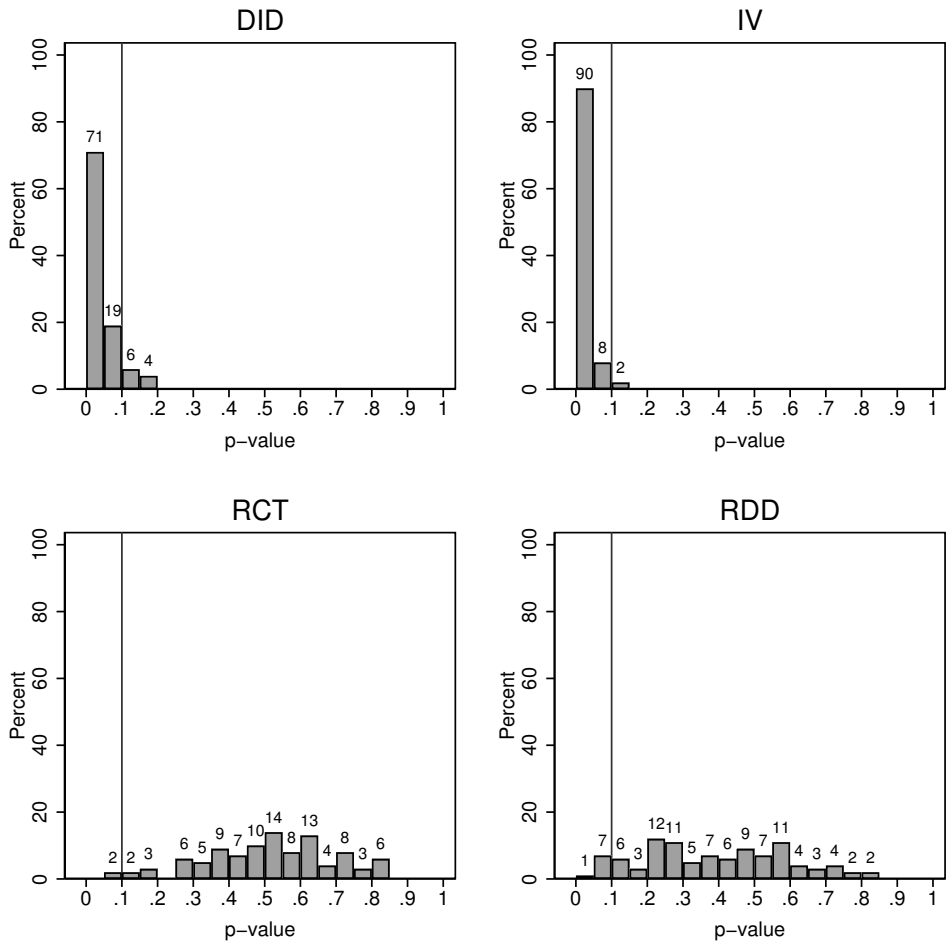

FIGURE A24. BOOTSTRAP RANDOMIZATION TESTS $z = 1.96$

Note: Tests have a null hypothesis that there is an equal proportion of tests just above and below $z = 1.96$. We use a window of half-width 0.25. From this set, we randomly select only one test from each table. By method, we then conduct the randomization test 100 times. The figures present histograms of the p-value. No weights have been applied.

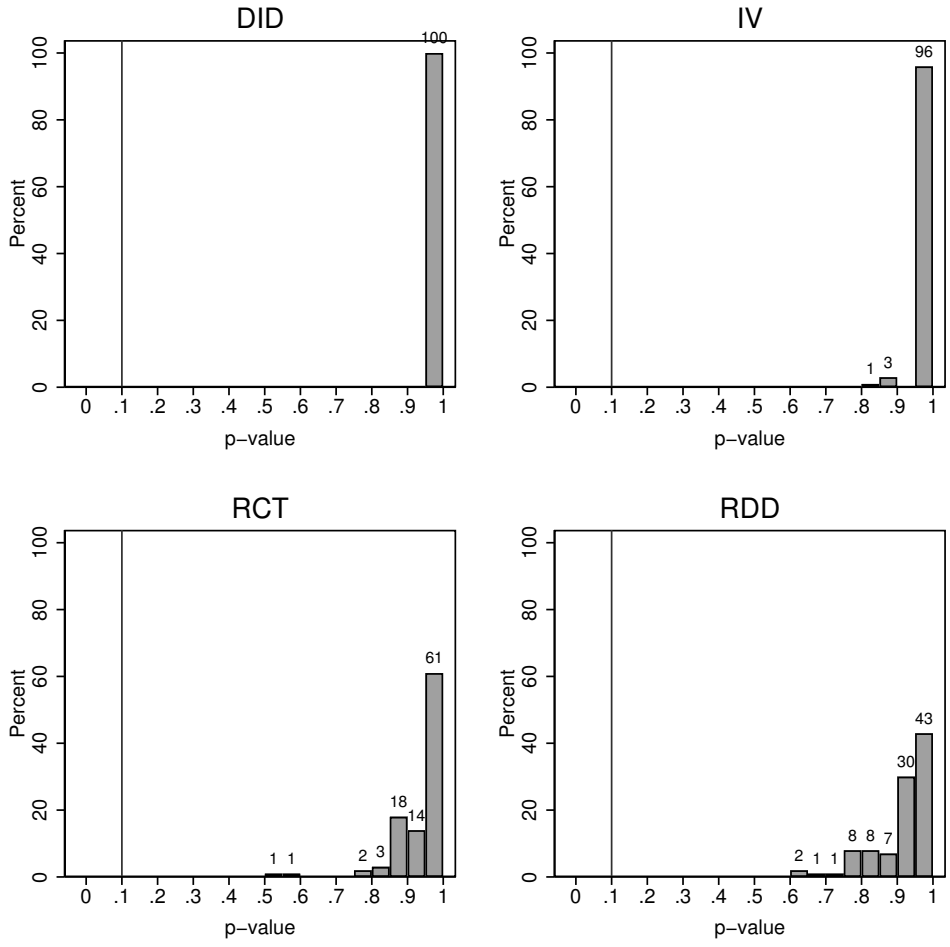

FIGURE A25. BOOTSTRAP RANDOMIZATION TESTS $z = 2.58$

Note: Tests have a null hypothesis that there is an equal proportion of tests just above and below $z = 2.58$. We use a window of half-width 0.25. From this set, we randomly select only one test from each table. By method, we then conduct the randomization test 100 times. The figures present histograms of the p-value. No weights have been applied.

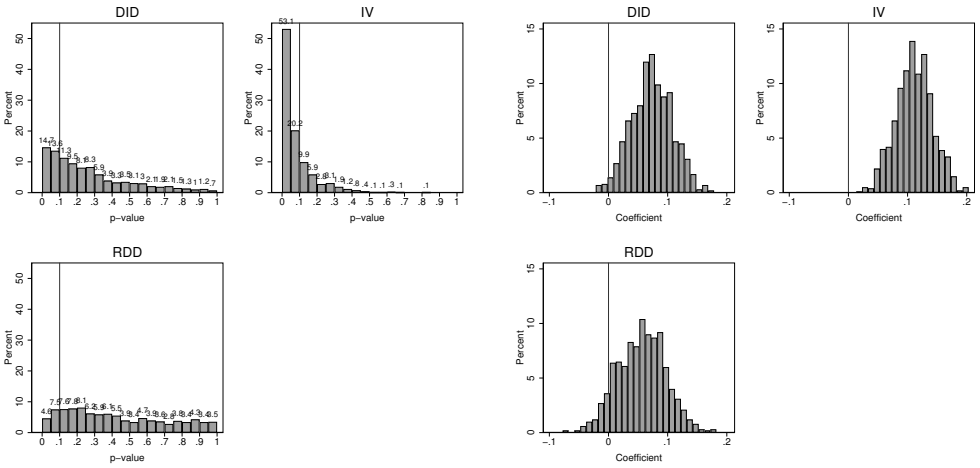

FIGURE A26. CALIPER TEST BOOTSTRAP FOR $z = 1.96$, RANDOM T STATISTICS WITHIN ALL PAPERS

Note: We randomly sample (50% of) the t statistics within each paper and conduct the same caliper tests as in the full sample. We present the results of bootstrapping the caliper tests by randomly sampling t statistics within papers 1,000 times. These would coincide with the coefficients from Appendix Table A22, in the first column (no controls). The left panel displays histograms of the p-values for each of the regression coefficients. The right panel displays histograms of the associated estimated effect sizes. The underlying caliper tests use article weights.

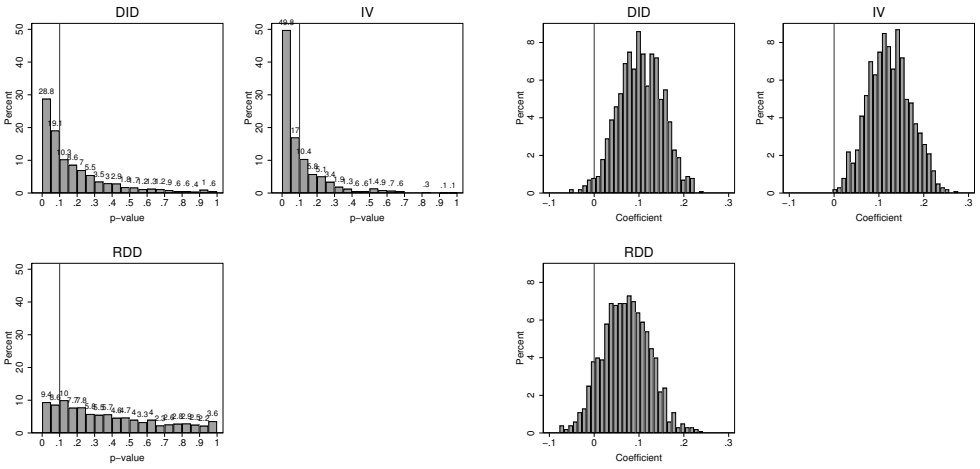

FIGURE A27. CALIPER TEST BOOTSTRAP FOR $z = 1.96$, RANDOM INCLUSION OF PAPERS

Note: We present the results of bootstrapping the caliper tests by randomly including papers in our sample 1,000 times. The underlying caliper tests use article weights.

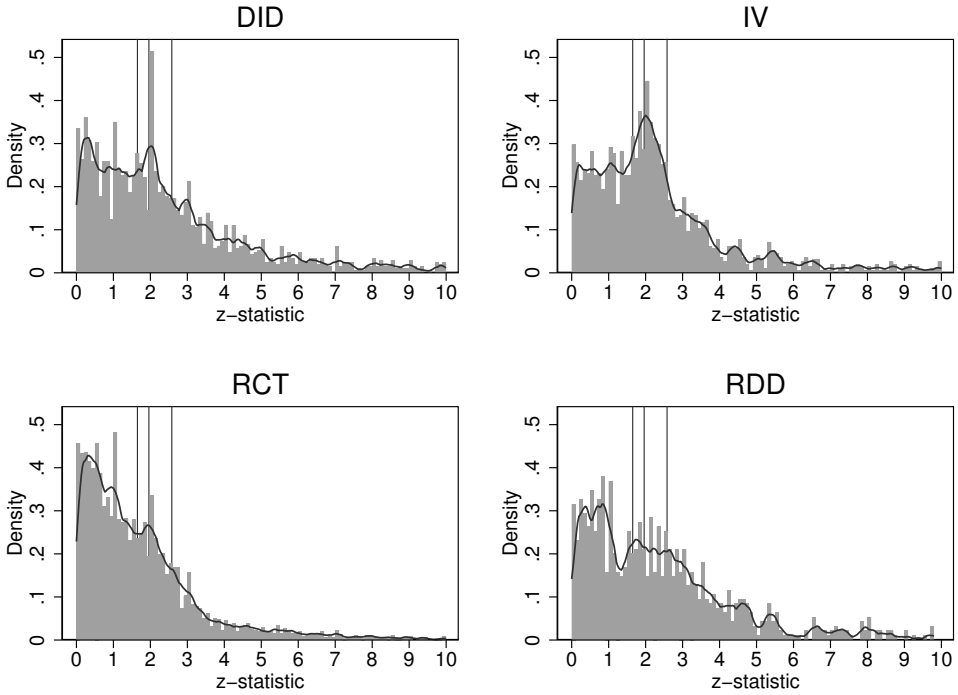

FIGURE A28. z-STATISTICS BY METHOD, AUTHORS WITH MULTIPLE JOURNAL ARTICLES

Note: This figure displays histograms of test statistics for $z \in [0, 10]$ by method: difference-in-differences (DID), instrumental variables (IV), randomized control trial (RCT), and regression discontinuity design (RDD). The sample is restricted to journal articles written by authors with at least two journal articles. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We have also superimposed an Epanechnikov kernel. No weights have been applied.

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	

Note: This table presents the 35 articles published in *American Economic Journal: Applied Economics* in 2015. We present the *American Economic Journal: Applied Economics* because it published work applying 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—ARTICLE AND AUTHOR CHARACTERISTICS

	(1) DID	(2) IV	(3) RCT	(4) RDD	(5) 2015	(6) 2018	(7) Top 5	(8) Non Top 5	(9) Total
Top 5	0.14	0.17	0.21	0.18	0.17	0.17	1.00	0.00	0.17
	(0.35)	(0.37)	(0.41)	(0.38)	(0.37)	(0.37)	(0.00)	(0.00)	(0.37)
Editor Present	0.63	0.62	0.66	0.58	0.65	0.61	0.74	0.60	0.63
	(0.48)	(0.49)	(0.47)	(0.50)	(0.48)	(0.49)	(0.44)	(0.49)	(0.48)
Solo-Authored	0.20	0.20	0.14	0.32	0.21	0.19	0.18	0.20	0.20
	(0.40)	(0.40)	(0.35)	(0.47)	(0.41)	(0.39)	(0.39)	(0.40)	(0.40)
Average Experience	10.31	11.00	11.00	8.76	10.67	10.42	11.03	10.42	10.52
	(5.68)	(6.61)	(5.52)	(5.17)	(6.23)	(5.82)	(6.33)	(5.93)	(6.00)
Female Authors	0.22	0.25	0.33	0.25	0.25	0.26	0.24	0.26	0.26
	(0.31)	(0.34)	(0.33)	(0.33)	(0.33)	(0.34)	(0.33)	(0.33)	(0.33)
Top Institutions	0.22	0.26	0.34	0.22	0.28	0.24	0.49	0.21	0.26
	(0.34)	(0.35)	(0.39)	(0.35)	(0.37)	(0.34)	(0.39)	(0.33)	(0.36)
Top PhD Institutions	0.37	0.35	0.48	0.31	0.31	0.43	0.56	0.34	0.38
	(0.39)	(0.39)	(0.39)	(0.38)	(0.37)	(0.40)	(0.39)	(0.38)	(0.39)
Articles	241	281	145	85	327	425	126	626	752

Note: Each observation is an article. (By test is presented in Table 2.) The Top 5 journals in economics are the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics* and *Review of Economic Studies*. Average experience is the mean of years since PhD for an article's authors. Share of female authors, share of authors affiliated with top institutions, and share of authors who completed a PhD at a top institution.

TABLE A3—RANDOMIZATION TESTS, 5% SIGNIFICANCE THRESHOLD, SAMPLING UNCERTAINTY

	(1) DID	(2) IV	(3) RCT	(4) RDD
Proportion Significant in 1.96±0.5	0.530	0.539	0.467	0.472
Standard Deviation	0.000	0.000	0.000	0.000
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	1365	1412	1719	706
Proportion Significant in 1.96±0.4	0.532	0.533	0.479	0.488
Standard Deviation	0.000	0.000	0.000	0.000
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	1137	1166	1416	582
Proportion Significant in 1.96±0.3	0.532	0.526	0.485	0.494
Standard Deviation	0.000	0.000	0.000	0.001
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	881	917	1098	453
Proportion Significant in 1.96±0.2	0.556	0.541	0.493	0.508
Standard Deviation	0.000	0.000	0.000	0.001
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	606	619	755	295
Proportion Significant in 1.96±0.1	0.631	0.575	0.547	0.542
Standard Deviation	0.001	0.001	0.001	0.002
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.043
Number of statistics in window	352	315	393	142
Proportion Significant in 1.96±0.075	0.684	0.597	0.560	0.565
Standard Deviation	0.001	0.001	0.001	0.002
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.104
Number of statistics in window	269	238	298	115
Proportion Significant in 1.96±0.05	0.707	0.601	0.641	0.614
Standard Deviation	0.001	0.001	0.001	0.003
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	208	168	209	83

Note: In this table we estimate the proportion p of statistical significance, by method and within a particular window. The p-value of equal probability refers to equal mass above and below the stated threshold. The p-value of equal to RCT refers to testing if the proportion of method i is significantly different from RCTs. In both, two sided tests are used. Articles are not weighted.

TABLE A4—RANDOMIZATION TESTS, 10% SIGNIFICANCE THRESHOLD, SAMPLING UNCERTAINTY

	(1) DID	(2) IV	(3) RCT	(4) RDD
Proportion Significant in 1.65±0.5	0.575	0.593	0.502	0.511
Standard Deviation	0.000	0.000	0.000	0.000
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	1310	1293	1876	734
Proportion Significant in 1.65±0.4	0.594	0.594	0.521	0.518
Standard Deviation	0.000	0.000	0.000	0.000
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	1056	1030	1498	589
Proportion Significant in 1.65±0.3	0.567	0.562	0.527	0.523
Standard Deviation	0.000	0.000	0.000	0.001
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	729	754	1084	440
Proportion Significant in 1.65±0.2	0.550	0.580	0.522	0.515
Standard Deviation	0.000	0.000	0.000	0.001
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	511	510	734	303
Proportion Significant in 1.65±0.1	0.510	0.567	0.513	0.537
Standard Deviation	0.001	0.001	0.001	0.002
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.049	0.000	.	0.000
Number of statistics in window	253	252	339	149
Proportion Significant in 1.65±0.075	0.522	0.568	0.517	0.578
Standard Deviation	0.001	0.001	0.001	0.002
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.030	0.000	.	0.000
Number of statistics in window	205	183	261	116
Proportion Significant in 1.65±0.05	0.490	0.596	0.512	0.560
Standard Deviation	0.002	0.002	0.002	0.003
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	149	136	164	75

Note: In this table we estimate the proportion p of statistical significance, by method and within a particular window. The p-value of equal probability refers to equal mass above and below the stated threshold. The p-value of equal to RCT refers to testing if the proportion of method i is significantly different from RCTs. In both, two sided tests are used. Articles are not weighted.

Table A5—Randomization Tests, 1% Significance Threshold, Sampling Uncertainty

	(1) DID	(2) IV	(3) RCT	(4) RDD
Proportion Significant in 2.58±0.5	0.429	0.400	0.397	0.399
Standard Deviation	0.000	0.000	0.000	0.000
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.001
Number of statistics in window	1106	1186	1146	536
Proportion Significant in 2.58±0.4	0.407	0.409	0.406	0.418
Standard Deviation	0.000	0.000	0.000	0.001
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.003	0.000	.	0.000
Number of statistics in window	859	930	892	407
Proportion Significant in 2.58±0.3	0.420	0.428	0.441	0.428
Standard Deviation	0.000	0.000	0.000	0.001
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	672	710	690	297
Proportion Significant in 2.58±0.2	0.424	0.452	0.484	0.415
Standard Deviation	0.001	0.001	0.001	0.001
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	434	484	461	195
Proportion Significant in 2.58±0.1	0.416	0.486	0.515	0.437
Standard Deviation	0.001	0.001	0.001	0.002
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	238	245	241	103
Proportion Significant in 2.58±0.075	0.448	0.521	0.555	0.431
Standard Deviation	0.002	0.001	0.002	0.003
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.000	.	0.000
Number of statistics in window	154	188	155	72
Proportion Significant in 2.58±0.05	0.398	0.515	0.513	0.419
Standard Deviation	0.002	0.002	0.002	0.006
p-value of equal probability	0.000	0.000	0.000	0.000
p-value of equal to RCT	0.000	0.610	.	0.000
Number of statistics in window	98	130	113	43

Note: In this table we estimate the proportion p of statistical significance, by method and within a particular window. The p-value of equal probability refers to equal mass above and below the stated threshold. The p-value of equal to RCT refers to testing if the proportion of method i is significantly different from RCTs. In both, two sided tests are used. Articles are not weighted.

TABLE A6—RANDOMIZATION TESTS, 10% THRESHOLD, UNWEIGHTED

	(1) DID	(2) IV	(3) RCT	(4) RDD
Proportion Significant in 1.65±0.5	0.575	0.593	0.502	0.511
One Sided p-value	0.000	0.000	0.454	0.290
Number of Tests in 1.65±0.5	1310	1293	1876	734
Proportion Significant in 1.65±0.4	0.594	0.594	0.521	0.518
One Sided p-value	0.000	0.000	0.057	0.205
Number of Tests in 1.65±0.4	1056	1030	1498	589
Proportion Significant in 1.65±0.3	0.567	0.562	0.527	0.523
One Sided p-value	0.000	0.000	0.042	0.183
Number of Tests in 1.65±0.3	729	754	1084	440
Proportion Significant in 1.65±0.2	0.550	0.580	0.522	0.515
One Sided p-value	0.013	0.000	0.126	0.323
Number of Tests in 1.65±0.2	511	510	734	303
Proportion Significant in 1.65±0.1	0.510	0.567	0.513	0.537
One Sided p-value	0.401	0.019	0.332	0.206
Number of Tests in 1.65±0.1	253	252	339	149
Proportion Significant in 1.65±0.075	0.522	0.568	0.517	0.578
One Sided p-value	0.288	0.038	0.310	0.057
Number of Tests in 1.65±0.075	205	183	261	116
Proportion Significant in 1.65±0.05	0.490	0.596	0.512	0.560
One Sided p-value	0.628	0.016	0.407	0.178
Number of Tests in 1.65±0.05	149	136	164	75

Note: In this table we present the results of binomial proportion tests where a success is defined as a statistically significant observation at the threshold level. In the first panel we use observations where $(1.15 < z < 2.15)$. The other panels use observations for smaller windows. 59.3% of the 1,293 IV tests within this window are significant. We then test if this proportion is statistically greater than 0.5. The associated p-values are then reported. Articles are not weighted.

Table A7—Randomization Tests, 1% Threshold, Unweighted

	(1) DID	(2) IV	(3) RCT	(4) RDD
Proportion Significant in 2.58±0.5	0.429	0.400	0.397	0.399
One Sided p-value	1.000	1.000	1.000	1.000
Number of Tests in 2.58±0.5	1106	1186	1146	536
Proportion Significant in 2.58±0.4	0.407	0.409	0.406	0.418
One Sided p-value	1.000	1.000	1.000	1.000
Number of Tests in 2.58±0.4	859	930	892	407
Proportion Significant in 2.58±0.3	0.420	0.428	0.441	0.428
One Sided p-value	1.000	1.000	0.999	0.995
Number of Tests in 2.58±0.3	672	710	690	297
Proportion Significant in 2.58±0.2	0.424	0.452	0.484	0.415
One Sided p-value	0.999	0.984	0.772	0.993
Number of Tests in 2.58±0.2	434	484	461	195
Proportion Significant in 2.58±0.1	0.416	0.486	0.515	0.437
One Sided p-value	0.996	0.695	0.350	0.916
Number of Tests in 2.58±0.1	238	245	241	103
Proportion Significant in 2.58±0.075	0.448	0.521	0.555	0.431
One Sided p-value	0.915	0.305	0.099	0.903
Number of Tests in 2.58±0.075	154	188	155	72
Proportion Significant in 2.58±0.05	0.398	0.515	0.513	0.419
One Sided p-value	0.983	0.396	0.425	0.889
Number of Tests in 2.58±0.05	98	130	113	43

Note: In this table we present the results of binomial proportion tests where a success is defined as a statistically significant observation at the threshold level. In the first panel we use observations where $(2.08 < z < 3.08)$. The other panels use observations for smaller windows. In the first panel, 40% of the 1,186 IV tests within this window are significant. We then test if this proportion is statistically greater than 0.5. The associated p-values are then reported. Articles are not weighted.

TABLE A8—RANDOMIZATION TESTS, 1% SIGNIFICANCE THRESHOLD, ARTICLE WEIGHTED

	(1) DID	(2) IV	(3) RCT	(4) RDD
Proportion Significant in 2.58±0.5	0.454	0.359	0.397	0.395
One Sided p-value	1.000	1.000	1.000	1.000
Number of Tests in 2.58±0.5	5.8e+04	5.0e+04	1.0e+05	3.5e+04
Proportion Significant in 2.58±0.4	0.404	0.371	0.395	0.402
One Sided p-value	1.000	1.000	1.000	1.000
Number of Tests in 2.58±0.4	4.4e+04	3.9e+04	7.7e+04	2.8e+04
Proportion Significant in 2.58±0.3	0.418	0.397	0.434	0.412
One Sided p-value	1.000	1.000	1.000	1.000
Number of Tests in 2.58±0.3	3.4e+04	3.0e+04	5.9e+04	2.0e+04
Proportion Significant in 2.58±0.2	0.451	0.419	0.464	0.391
One Sided p-value	1.000	1.000	1.000	1.000
Number of Tests in 2.58±0.2	2.1e+04	2.0e+04	4.0e+04	1.3e+04
Proportion Significant in 2.58±0.1	0.427	0.409	0.492	0.409
One Sided p-value	1.000	1.000	0.991	1.000
Number of Tests in 2.58±0.1	1.2e+04	9956.000	2.0e+04	7165.000
Proportion Significant in 2.58±0.075	0.441	0.426	0.559	0.383
One Sided p-value	1.000	1.000	0.000	1.000
Number of Tests in 2.58±0.075	7155.000	8052.000	1.2e+04	5174.000
Proportion Significant in 2.58±0.05	0.374	0.421	0.551	0.368
One Sided p-value	1.000	1.000	0.000	1.000
Number of Tests in 2.58±0.05	4318.000	5325.000	9237.000	3235.000

Note: In this table we present the results of binomial proportion tests where a success is defined as a statistically significant observation at the threshold level. Successive panels use observations from smaller windows. We then test if this proportion is statistically greater than 0.5. The associated p-values are then reported. Weighted by the number of tests in an article.

Table A9—Randomization Tests, 5% Significance Threshold, Article Weighted

	(1) DID	(2) IV	(3) RCT	(4) RDD
Proportion Significant in 1.96±0.5	0.521	0.514	0.456	0.437
One Sided p-value	0.000	0.000	1.000	1.000
Number of Tests in 1.96±0.5	7.5e+04	6.5e+04	1.6e+05	5.1e+04
Proportion Significant in 1.96±0.4	0.538	0.512	0.474	0.454
One Sided p-value	0.000	0.000	1.000	1.000
Number of Tests in 1.96±0.4	6.3e+04	5.2e+04	1.3e+05	4.2e+04
Proportion Significant in 1.96±0.3	0.545	0.505	0.496	0.468
One Sided p-value	0.000	0.015	0.998	1.000
Number of Tests in 1.96±0.3	4.7e+04	4.0e+04	1.0e+05	3.2e+04
Proportion Significant in 1.96±0.2	0.576	0.524	0.502	0.479
One Sided p-value	0.000	0.000	0.099	1.000
Number of Tests in 1.96±0.2	3.3e+04	2.6e+04	7.2e+04	2.0e+04
Proportion Significant in 1.96±0.1	0.684	0.517	0.593	0.527
One Sided p-value	0.000	0.000	0.000	0.000
Number of Tests in 1.96±0.1	2.0e+04	1.3e+04	3.8e+04	1.0e+04
Proportion Significant in 1.96±0.075	0.738	0.525	0.611	0.547
One Sided p-value	0.000	0.000	0.000	0.000
Number of Tests in 1.96±0.075	1.7e+04	1.1e+04	2.9e+04	8144.000
Proportion Significant in 1.96±0.05	0.766	0.506	0.714	0.589
One Sided p-value	0.000	0.164	0.000	0.000
Number of Tests in 1.96±0.05	1.4e+04	7556.000	2.0e+04	5496.000

Note: In this table we present the results of binomial proportion tests where a success is defined as a statistically significant observation at the threshold level. Successive panels use observations from smaller windows. We then test if this proportion is statistically greater than 0.5. The associated p-values are then reported. Weighted by the number of tests in an article.

TABLE A10—RANDOMIZATION TESTS, 10% SIGNIFICANCE THRESHOLD, ARTICLE WEIGHTED

	(1) DID	(2) IV	(3) RCT	(4) RDD
Proportion Significant in 1.65±0.5	0.534	0.539	0.471	0.490
One Sided p-value	0.000	0.000	1.000	1.000
Number of Tests in 1.65±0.5	7.6e+04	6.2e+04	1.9e+05	5.5e+04
Proportion Significant in 1.65±0.4	0.568	0.539	0.493	0.499
One Sided p-value	0.000	0.000	1.000	0.664
Number of Tests in 1.65±0.4	6.2e+04	5.0e+04	1.5e+05	4.5e+04
Proportion Significant in 1.65±0.3	0.518	0.528	0.492	0.506
One Sided p-value	0.000	0.000	1.000	0.017
Number of Tests in 1.65±0.3	4.1e+04	3.6e+04	1.1e+05	3.4e+04
Proportion Significant in 1.65±0.2	0.507	0.541	0.497	0.501
One Sided p-value	0.005	0.000	0.961	0.427
Number of Tests in 1.65±0.2	3.0e+04	2.5e+04	7.3e+04	2.3e+04
Proportion Significant in 1.65±0.1	0.458	0.556	0.477	0.520
One Sided p-value	1.000	0.000	1.000	0.000
Number of Tests in 1.65±0.1	1.4e+04	1.3e+04	3.2e+04	1.2e+04
Proportion Significant in 1.65±0.075	0.452	0.508	0.481	0.560
One Sided p-value	1.000	0.068	1.000	0.000
Number of Tests in 1.65±0.075	1.2e+04	8583.000	2.4e+04	9022.000
Proportion Significant in 1.65±0.05	0.418	0.563	0.448	0.535
One Sided p-value	1.000	0.000	1.000	0.000
Number of Tests in 1.65±0.05	8536.000	6575.000	1.5e+04	5708.000

Note: In this table we present the results of binomial proportion tests where a success is defined as a statistically significant observation at the threshold level. Successive panels use observations from smaller windows. We then test if this proportion is statistically greater than 0.5. The associated p-values are then reported. Weighted by the number of tests in an article.

Table A11—Randomization Tests for Other Literatures

Economics	Total	Over	Under	Ratio	p Value
Combined	2275	1194	1081	1.10	0.009
DID	606	337	269	1.25	0.003
IV	619	335	284	1.18	0.022
RCT	755	372	383	0.97	0.668
RDD	295	150	145	1.03	0.408
Top 5	485	248	237	1.05	0.325
Non Top 5	1790	946	844	1.12	0.008
Political Science					
Combined*	192	139	53	2.62	<0.001
APSR	64	49	15	3.27	<0.001
AJPS	128	90	38	2.37	<0.001
Sociology					
Combined	106	73	33	2.21	<0.001
AJR + AJS*	77	51	26	1.96	0.003
ASR*	41	26	15	1.73	0.059
AJS*	36	25	11	2.27	0.014
TSQ	29	22	7	3.14	0.004

Note: Authors' calculations. All p-values from one-sided binomial tests. All tests use $1.76 < z < 2.16$. Economics from this manuscript using 25 top journals in Economics for 2015 and 2018. Political Science from **?** using *American Political Science Review* (APSR) and the *American Journal of Political Science* (AJPS) for 1995–2007. Sociology from **?** using *American Sociological Review* (ASR), the *American Journal of Sociology* (AJS), and the *Sociological Quarterly* (TSQ) for 2003–2005. ASR and AJS separated as they are widely considered the top two journals in sociology **?**.

TABLE A12—CALIPER TEST, SIGNIFICANT AT THE 5% LEVEL: LOGIT

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.095	0.088	0.054	0.050	0.051	0.026
	(0.034)	(0.033)	(0.032)	(0.033)	(0.037)	(0.047)
IV	0.102	0.097	0.072	0.080	0.091	0.088
	(0.034)	(0.034)	(0.033)	(0.033)	(0.037)	(0.045)
RDD	0.058	0.057	0.025	0.016	0.025	0.012
	(0.047)	(0.048)	(0.045)	(0.046)	(0.049)	(0.054)
Top 5		-0.051	-0.010			
		(0.045)	(0.084)			
Year=2018		0.020	0.030	0.024	0.010	0.043
		(0.028)	(0.027)	(0.027)	(0.030)	(0.035)
Experience		-0.002	-0.006	-0.005	-0.005	0.009
		(0.007)	(0.007)	(0.007)	(0.008)	(0.009)
Experience^2		-0.005	0.005	0.006	0.013	-0.029
		(0.018)	(0.018)	(0.019)	(0.020)	(0.025)
Top Institution		0.019	0.027	0.026	0.001	-0.004
		(0.050)	(0.044)	(0.043)	(0.046)	(0.055)
PhD Top Institution		-0.011	-0.031	-0.023	0.022	0.067
		(0.039)	(0.037)	(0.038)	(0.040)	(0.048)
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	5,202	5,202	5,202	5,202	3,798	2,273
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]
RCT Sig Rate	0.47	0.47	0.47	0.47	0.48	0.49

Note: This table reports marginal effects from logit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. In columns 1–4, we restrict the sample to $z \in [1.46, 2.46]$. Column 5 restricts the sample to $z \in [1.61, 2.31]$, while columns 6 restricts the sample to $z \in [1.76, 2.16]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

TABLE A13—CALIPER TEST, SIGNIFICANT AT THE 5% LEVEL, BOOTSTRAP ERRORS: UNWEIGHTED ESTIMATES

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.062	0.056	0.037	0.041	0.033	0.063
	(0.023)	(0.024)	(0.025)	(0.025)	(0.033)	(0.043)
IV	0.072	0.059	0.048	0.054	0.040	0.061
	(0.023)	(0.025)	(0.027)	(0.028)	(0.035)	(0.035)
RDD	0.005	-0.002	-0.025	-0.019	-0.004	0.016
	(0.030)	(0.030)	(0.034)	(0.038)	(0.044)	(0.037)
Top 5		-0.032	0.008			
		(0.025)	(0.080)			
Year=2018		0.024	0.030	0.028	0.019	0.030
		(0.021)	(0.021)	(0.022)	(0.027)	(0.029)
Experience		0.001	0.001	0.001	0.000	0.018
		(0.006)	(0.006)	(0.006)	(0.007)	(0.009)
Experience^2		-0.009	-0.008	-0.006	-0.003	-0.055
		(0.020)	(0.020)	(0.021)	(0.025)	(0.031)
Top Institution		-0.011	-0.014	-0.010	0.012	-0.008
		(0.034)	(0.037)	(0.035)	(0.045)	(0.043)
PhD Top Institution		-0.004	-0.013	-0.024	-0.023	0.022
		(0.031)	(0.036)	(0.034)	(0.040)	(0.041)
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	5,202	5,202	5,202	5,202	3,798	2,273
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]
RCT Sig Rate	0.47	0.47	0.47	0.47	0.48	0.49

Note: This table reports marginal effects from probit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. In columns 1–4, we restrict the sample to $z \in [1.46, 2.46]$. Column 5 restricts the sample to $z \in [1.61, 2.31]$, while columns 6 restricts the sample to $z \in [1.76, 2.16]$. Bootstrapped standard errors are in parentheses, resampled by article 250 times.

TABLE A14—CALIPER TEST, SIGNIFICANT AT THE 5% LEVEL: UNWEIGHTED ESTIMATES

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.062	0.056	0.037	0.041	0.033	0.063
	(0.025)	(0.024)	(0.025)	(0.026)	(0.030)	(0.036)
IV	0.072	0.059	0.048	0.054	0.040	0.061
	(0.023)	(0.025)	(0.027)	(0.028)	(0.030)	(0.036)
RDD	0.005	-0.002	-0.025	-0.019	-0.004	0.016
	(0.029)	(0.031)	(0.033)	(0.034)	(0.037)	(0.042)
Top 5		-0.032	0.008			
		(0.028)	(0.071)			
Year=2018		0.024	0.030	0.028	0.019	0.030
		(0.021)	(0.021)	(0.021)	(0.023)	(0.027)
Experience		0.001	0.001	0.001	0.000	0.018
		(0.006)	(0.006)	(0.006)	(0.006)	(0.008)
Experience^2		-0.009	-0.008	-0.006	-0.003	-0.055
		(0.020)	(0.018)	(0.018)	(0.021)	(0.027)
Top Institution		-0.011	-0.014	-0.010	0.012	-0.008
		(0.033)	(0.034)	(0.033)	(0.036)	(0.042)
PhD Top Institution		-0.004	-0.013	-0.024	-0.023	0.022
		(0.031)	(0.032)	(0.031)	(0.035)	(0.040)
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	5,202	5,202	5,202	5,202	3,798	2,273
RCT Sig Rate	0.47	0.47	0.47	0.47	0.48	0.49

Note: This table reports marginal effects from probit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. In columns 1–4, we restrict the sample to $z \in [1.46, 2.46]$. Column 5 restricts the sample to $z \in [1.61, 2.31]$, while columns 6 restricts the sample to $z \in [1.76, 2.16]$. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table A15—Significant at the 5% Level, Expanded Covariates

	(1)	(2)	(3)	(4)	(5)	(6)
Top 5		-0.051	-0.010			
		(0.045)	(0.084)			
Year=2018		0.021	0.030	0.024	0.010	0.043
		(0.028)	(0.027)	(0.027)	(0.030)	(0.035)
Experience		-0.002	-0.006	-0.005	-0.006	0.009
		(0.007)	(0.007)	(0.007)	(0.008)	(0.009)
Experience^2		-0.005	0.005	0.006	0.014	-0.028
		(0.018)	(0.018)	(0.019)	(0.020)	(0.025)
Top Institution		0.019	0.026	0.025	-0.001	-0.005
		(0.050)	(0.044)	(0.043)	(0.046)	(0.055)
PhD Top Institution		-0.011	-0.030	-0.023	0.023	0.067
		(0.039)	(0.037)	(0.038)	(0.040)	(0.048)
Reported as p-value		-0.139	-0.172	-0.033	0.075	
		(0.109)	(0.103)	(0.181)	(0.169)	
Reported as coeff.		-0.057	-0.063	0.068	0.179	
		(0.086)	(0.085)	(0.172)	(0.157)	
Reported as t-stat.		-0.031	-0.121	0.005	0.182	
		(0.099)	(0.104)	(0.184)	(0.169)	
Solo-Authored		-0.044	-0.062	-0.048	-0.030	-0.017
		(0.040)	(0.039)	(0.039)	(0.040)	(0.049)
Share Female Authors		0.002	0.022	0.021	0.027	0.014
		(0.040)	(0.038)	(0.037)	(0.040)	(0.046)
Editor Present		0.008	0.008	0.012	0.019	-0.043
		(0.036)	(0.035)	(0.036)	(0.036)	(0.046)
Finance			0.151			
			(0.085)			
Macroeconomics			0.080			
			(0.099)			
Gen. Int. (Non Top 5)			0.042			
			(0.077)			
Experimental			-0.251			
			(0.178)			
Development			-0.025			
			(0.084)			
Labor			0.052			
			(0.089)			
Public			0.008			
			(0.078)			
Urban			0.003			
			(0.094)			
Journal FE				Y	Y	Y
Observations	5,202	5,202	5,202	5,202	3,798	2,273
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]
RCT Sig Rate	0.47	0.47	0.47	0.47	0.48	0.49

Note: This table reports marginal effects from probit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

TABLE A16—CALIPER TEST, SIGNIFICANT AT THE 10% LEVEL

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.043	0.052	0.057	0.062	0.030	0.055
	(0.037)	(0.036)	(0.035)	(0.037)	(0.042)	(0.051)
IV	0.089	0.078	0.074	0.074	0.030	0.043
	(0.035)	(0.035)	(0.036)	(0.036)	(0.038)	(0.047)
RDD	-0.028	-0.043	-0.032	-0.019	-0.035	-0.032
	(0.034)	(0.034)	(0.036)	(0.037)	(0.044)	(0.051)
Top 5		0.050	-0.064			
		(0.039)	(0.092)			
Year=2018		0.001	-0.002	-0.003	0.008	0.038
		(0.029)	(0.028)	(0.029)	(0.031)	(0.035)
Experience		-0.000	-0.002	-0.003	-0.009	0.001
		(0.007)	(0.007)	(0.007)	(0.007)	(0.008)
Experience^2		-0.010	-0.003	-0.001	0.015	-0.010
		(0.021)	(0.021)	(0.021)	(0.022)	(0.022)
Top Institution		-0.012	-0.018	-0.010	0.013	0.032
		(0.048)	(0.041)	(0.041)	(0.041)	(0.051)
PhD Top Institution		-0.077	-0.070	-0.065	-0.068	-0.130
		(0.040)	(0.037)	(0.036)	(0.039)	(0.051)
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	5,213	5,213	5,213	5,213	3,477	2,053
Window	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]	[1.65±0.35]	[1.65±0.20]
RCT Sig Rate	0.50	0.50	0.50	0.50	0.50	0.52

Note: This table reports marginal effects from probit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. In columns 1–4, we restrict the sample to $z \in [1.30, 2.00]$. Column 5 restricts the sample to $z \in [1.61, 2.31]$, while columns 6 restricts the sample to $z \in [1.45, 1.85]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

TABLE A17—CALIPER TEST, SIGNIFICANT AT THE 10% LEVEL: UNWEIGHTED ESTIMATES

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.074	0.052	0.040	0.041	0.001	0.015
	(0.025)	(0.024)	(0.025)	(0.025)	(0.029)	(0.037)
IV	0.092	0.065	0.055	0.056	0.029	0.031
	(0.025)	(0.025)	(0.027)	(0.028)	(0.030)	(0.035)
RDD	0.010	-0.027	-0.046	-0.046	-0.054	-0.023
	(0.025)	(0.026)	(0.029)	(0.029)	(0.034)	(0.043)
Top 5		0.039	0.004			
		(0.027)	(0.052)			
Year=2018		-0.017	-0.010	-0.016	-0.021	-0.017
		(0.018)	(0.018)	(0.018)	(0.021)	(0.026)
Experience		-0.005	-0.005	-0.005	-0.009	-0.002
		(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
Experience^2		0.004	0.004	0.004	0.019	0.005
		(0.020)	(0.019)	(0.019)	(0.022)	(0.025)
Top Institution		-0.078	-0.072	-0.067	-0.057	-0.096
		(0.029)	(0.028)	(0.028)	(0.031)	(0.040)
PhD Top Institution		-0.016	-0.024	-0.024	-0.021	-0.034
		(0.025)	(0.026)	(0.027)	(0.031)	(0.037)
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	5,213	5,213	5,213	5,213	3,477	2,053
RCT Sig Rate	0.50	0.50	0.50	0.50	0.50	0.52

Note: This table reports marginal effects from probit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. In columns 1–4, we restrict the sample to $z \in [1.30, 2.00]$. Column 5 restricts the sample to $z \in [1.61, 2.31]$, while columns 6 restricts the sample to $z \in [1.45, 1.85]$. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

TABLE A18—CALIPER TEST, SIGNIFICANT AT THE 1% LEVEL

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.037	0.027	0.001	-0.005	0.028	-0.073
	(0.041)	(0.038)	(0.040)	(0.039)	(0.045)	(0.052)
IV	0.017	0.003	-0.025	-0.025	0.013	-0.015
	(0.038)	(0.037)	(0.037)	(0.037)	(0.041)	(0.049)
RDD	-0.001	0.004	-0.019	-0.027	0.026	-0.061
	(0.066)	(0.062)	(0.057)	(0.050)	(0.057)	(0.072)
Top 5		0.035	-0.126			
		(0.044)	(0.113)			
Year=2018		0.040	0.039	0.042	0.064	0.069
		(0.035)	(0.035)	(0.034)	(0.036)	(0.040)
Experience		0.007	0.004	0.005	0.006	0.007
		(0.009)	(0.008)	(0.008)	(0.010)	(0.010)
Experience^2		-0.009	-0.001	-0.007	-0.011	-0.039
		(0.029)	(0.027)	(0.029)	(0.038)	(0.031)
Top Institution		-0.009	0.005	0.005	-0.007	0.064
		(0.050)	(0.050)	(0.049)	(0.051)	(0.062)
PhD Top Institution		-0.078	-0.079	-0.060	-0.022	-0.076
		(0.044)	(0.044)	(0.043)	(0.048)	(0.056)
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	3,974	3,974	3,974	3,974	2,722	1,570
Window	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]	[2.58±0.35]	[2.58±0.20]
RCT Sig Rate	0.40	0.40	0.40	0.40	0.42	0.48

Note: This table reports marginal effects from probit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. In columns 1–4, we restrict the sample to $z \in [2.08, 3.08]$. Column 5 restricts the sample to $z \in [2.23, 2.93]$, while columns 6 restricts the sample to $z \in [2.38, 2.78]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

TABLE A19—CALIPER TEST, SIGNIFICANT AT THE 1% LEVEL: UNWEIGHTED ESTIMATES

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.031	0.036	0.031	0.030	0.012	-0.035
	(0.026)	(0.027)	(0.028)	(0.029)	(0.032)	(0.044)
IV	0.003	0.005	-0.001	0.002	0.017	0.008
	(0.025)	(0.025)	(0.026)	(0.025)	(0.029)	(0.037)
RDD	0.002	0.004	0.000	0.006	0.019	-0.048
	(0.031)	(0.030)	(0.031)	(0.033)	(0.036)	(0.048)
Top 5		0.087	0.037			
		(0.029)	(0.067)			
Year=2018		0.030	0.039	0.034	0.065	0.052
		(0.022)	(0.022)	(0.022)	(0.023)	(0.030)
Experience		0.002	0.001	0.002	-0.000	-0.005
		(0.005)	(0.005)	(0.005)	(0.006)	(0.008)
Experience^2		-0.009	-0.006	-0.013	-0.007	-0.011
		(0.015)	(0.015)	(0.015)	(0.020)	(0.028)
Top Institution		-0.024	-0.014	-0.023	0.002	0.024
		(0.031)	(0.031)	(0.031)	(0.033)	(0.043)
PhD Top Institution		-0.023	-0.036	-0.032	-0.053	-0.067
		(0.032)	(0.030)	(0.030)	(0.031)	(0.045)
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	3,974	3,974	3,974	3,974	2,722	1,570
RCT Sig Rate	0.40	0.40	0.40	0.40	0.42	0.48

Note: This table reports marginal effects from probit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. In columns 1–4, we restrict the sample to $z \in [2.08, 3.08]$. Column 5 restricts the sample to $z \in [2.23, 2.93]$, while columns 6 restricts the sample to $z \in [2.38, 2.78]$. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

TABLE A20—CALIPER TEST, SIGNIFICANT AT THE 5% LEVEL: WINDOWS

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.102	0.093	0.088	0.091	0.097	0.186
	(0.032)	(0.033)	(0.037)	(0.042)	(0.050)	(0.071)
IV	0.093	0.098	0.096	0.090	0.124	0.197
	(0.032)	(0.034)	(0.037)	(0.042)	(0.053)	(0.068)
RDD	0.060	0.059	0.047	0.056	0.060	0.092
	(0.047)	(0.048)	(0.049)	(0.048)	(0.059)	(0.072)
Year=2018	0.020	0.021	0.014	0.015	0.038	0.024
	(0.026)	(0.028)	(0.029)	(0.032)	(0.039)	(0.053)
Experience	-0.004	-0.002	-0.004	-0.001	0.009	0.001
	(0.007)	(0.007)	(0.008)	(0.008)	(0.010)	(0.013)
Experience^2	0.004	-0.005	0.003	0.000	-0.033	-0.004
	(0.019)	(0.018)	(0.021)	(0.021)	(0.026)	(0.037)
Top Institution	-0.013	0.009	-0.005	-0.045	-0.050	-0.045
	(0.041)	(0.045)	(0.049)	(0.053)	(0.069)	(0.094)
PhD Top Institution	0.013	-0.017	-0.006	0.037	0.073	0.039
	(0.038)	(0.041)	(0.045)	(0.047)	(0.061)	(0.081)
Reporting Method	Y	Y	Y	Y	Y	Y
Solo Authored	Y	Y	Y	Y	Y	Y
Share Female Authors	Y	Y	Y	Y	Y	Y
Editor	Y	Y	Y	Y	Y	Y
Observations	6,070	5,202	4,301	3,349	2,273	1,201
Window	[1.96±0.60]	[1.96±0.50]	[1.96±0.40]	[1.96±0.30]	[1.96±0.20]	[1.96±0.10]
RCT Sig Rate	0.46	0.47	0.48	0.49	0.49	0.55

Note: This table reports marginal effects from probit regressions (Equation (2)). 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 $z \in [1.46, 2.46]$. In column 3, we restrict the sample to $z \in [1.56, 2.36]$. Column 4 restricts the sample to $z \in [1.66, 2.26]$. In column 5, we restrict the sample to $z \in [1.76, 2.16]$. Column 6 restricts the sample to $z \in [1.86, 2.16]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

TABLE A21—CALIPER TEST, SIGNIFICANT AT THE 5% LEVEL: WINDOWS WITH JOURNAL FIXED EFFECTS

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.062	0.051	0.051	0.053	0.027	0.061
	(0.033)	(0.033)	(0.036)	(0.040)	(0.047)	(0.060)
IV	0.070	0.080	0.086	0.084	0.089	0.132
	(0.033)	(0.033)	(0.037)	(0.039)	(0.045)	(0.057)
RDD	0.023	0.016	0.012	0.029	0.012	0.013
	(0.045)	(0.046)	(0.048)	(0.048)	(0.055)	(0.069)
Year=2018	0.024	0.024	0.007	0.012	0.043	0.033
	(0.026)	(0.027)	(0.029)	(0.031)	(0.035)	(0.042)
Experience	-0.009	-0.005	-0.006	-0.001	0.009	0.006
	(0.007)	(0.007)	(0.008)	(0.008)	(0.009)	(0.011)
Experience^2	0.018	0.006	0.011	0.004	-0.028	-0.012
	(0.019)	(0.019)	(0.021)	(0.021)	(0.025)	(0.037)
Top Institution	0.006	0.025	0.017	-0.012	-0.005	0.071
	(0.041)	(0.043)	(0.047)	(0.048)	(0.055)	(0.064)
PhD Top Institution	0.004	-0.023	-0.006	0.034	0.067	0.023
	(0.036)	(0.038)	(0.040)	(0.041)	(0.048)	(0.059)
Reporting Method	Y	Y	Y	Y	Y	Y
Solo Authored	Y	Y	Y	Y	Y	Y
Share Female Authors	Y	Y	Y	Y	Y	Y
Editor	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y
Observations	6,070	5,202	4,301	3,349	2,273	1,200
Window	[1.96±0.60]	[1.96±0.50]	[1.96±0.40]	[1.96±0.30]	[1.96±0.20]	[1.96±0.10]
RCT Sig Rate	0.46	0.47	0.48	0.49	0.49	0.55

Note: This table reports marginal effects from probit regressions (Equation (2)). 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 $z \in [1.46, 2.46]$. In column 3, we restrict the sample to $z \in [1.56, 2.36]$. Column 4 restricts the sample to $z \in [1.66, 2.26]$. In column 5, we restrict the sample to $z \in [1.76, 2.16]$. Column 6 restricts the sample to $z \in [1.86, 2.16]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

TABLE A22—CALIPER TEST, SIGNIFICANT AT THE 5% LEVEL: FIRST TABLE

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.061	0.084	0.090	0.083	0.027	-0.005
	(0.072)	(0.066)	(0.065)	(.)	(.)	(0.087)
IV	0.148	0.157	0.197	0.199	0.166	0.077
	(0.066)	(0.058)	(0.059)	(.)	(.)	(0.075)
RDD	0.032	0.056	0.055	0.040	0.040	0.093
	(0.074)	(0.078)	(0.080)	(.)	(.)	(0.111)
Top 5		0.042	0.205			
		(0.068)	(0.100)			
Year=2018		0.039	0.051	0.042	-0.057	0.005
		(0.053)	(0.051)	(.)	(.)	(0.057)
Experience		-0.012	-0.016	-0.007	-0.022	0.003
		(0.015)	(0.014)	(.)	(.)	(0.016)
Experience^2		0.029	0.030	0.007	0.062	-0.019
		(0.049)	(0.045)	(.)	(.)	(0.050)
Top Institution		-0.025	-0.061	-0.033	-0.026	0.185
		(0.081)	(0.080)	(.)	(.)	(0.111)
PhD Top Institution		-0.015	-0.029	-0.012	-0.007	-0.069
		(0.068)	(0.067)	(.)	(.)	(0.092)
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	1,566	1,566	1,566	1,566	1,139	687
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]
RCT Sig Rate	0.46	0.46	0.46	0.46	0.46	0.45

Note: This table reports marginal effects from probit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. In columns 1–4, we restrict the sample to $z \in [1.46, 2.46]$. Column 5 restricts the sample to $z \in [1.61, 2.31]$, while columns 6 restricts the sample to $z \in [1.76, 2.16]$. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A23—Excess Coefficients by Significance Region in Comparison to RCT

	(1) DID	(2) IV	(3) RDD
[0,1.65)			
Observed	0.360	0.338	0.465
Expected (RCT)	0.536	0.536	0.536
Difference	-0.176	-0.198	-0.071
Ratio of Excess to Expected	-0.328	-0.369	-0.132
[1.65,1.96)			
Observed	0.073	0.086	0.075
RCT	0.077	0.077	0.077
Difference	-0.005	0.009	-0.003
Ratio of Excess to Expected	-0.061	0.116	-0.034
[1.96,2.58)			
Observed	0.150	0.178	0.129
RCT	0.124	0.124	0.124
Difference	0.026	0.054	0.005
Ratio of Excess to Expected	0.211	0.433	0.041
[2.58,5)			
Observed	0.250	0.259	0.200
RCT	0.154	0.154	0.154
Difference	0.096	0.104	0.045
Ratio of Excess to Expected	0.621	0.677	0.295
[5,∞)			
Observed	0.153	0.136	0.122
RCT	0.094	0.094	0.094
Difference	0.059	0.042	0.028
Ratio of Excess to Expected	0.622	0.444	0.291

Note: Each panel of the table is a separate significance region. In each panel and for each method, we report four statistics. 1) The observed mass of test statistics 2) The expected mass informed by a calibrated t distribution 3) The difference and 4) The ratio of the observed to expected. For the difference and ratio, a negative value implies 'missing' test statistics in the region whereas a positive number implies an excess of test statistics. The degrees of freedom and the noncentrality parameter for the t distributions that fit the observed data best are presented at the bottom.

TABLE A24—EXCESS COEFFICIENTS BY SIGNIFICANCE REGION

	(1)	(2)	(3)	(4)
[0,1.65)	DID	IV	RCT	RDD
Observed	0.360	0.338	0.536	0.465
Expected	0.360	0.401	0.528	0.438
Difference	0.001	-0.063	0.008	0.027
Ratio of Excess to Expected	0.002	-0.157	0.015	0.062
[1.65,1.96)	DID	IV	RCT	RDD
Observed	0.073	0.086	0.077	0.075
Expected	0.086	0.086	0.079	0.085
Difference	-0.013	0.001	-0.002	-0.010
Ratio of Excess to Expected	-0.155	0.008	-0.027	-0.117
[1.96,2.58)	DID	IV	RCT	RDD
Observed	0.150	0.178	0.124	0.129
Expected	0.142	0.137	0.117	0.132
Difference	0.008	0.041	0.008	-0.002
Ratio of Excess to Expected	0.060	0.300	0.065	-0.018
[2.58,5)	DID	IV	RCT	RDD
Observed	0.250	0.259	0.154	0.200
Expected	0.260	0.240	0.182	0.223
Difference	-0.010	0.019	-0.028	-0.023
Ratio of Excess to Expected	-0.037	0.078	-0.153	-0.104
[5,∞)	DID	IV	RCT	RDD
Observed	0.153	0.136	0.094	0.122
Expected	0.153	0.136	0.094	0.123
Difference	0.001	0.000	0.000	-0.001
Ratio of Excess to Expected	0.004	0.001	0.004	-0.007
Degrees of Freedom	2	2	2	2
Non-centrality Parameter	1.81	1.65	1.16	1.51

Note: Each panel of the table is a separate significance region. In each panel and for each method, we report four statistics: 1) The observed mass of test statistics 2) The expected mass informed by a calibrated t distribution 3) The difference and 4) The ratio of the observed to expected. For the difference and ratio, a negative value implies 'missing' test statistics in the region whereas a positive number implies an excess of test statistics. The degrees of freedom and the noncentrality parameter for the t distributions that fit the observed data best are presented at the bottom.

TABLE A25—RANDOMIZATION TESTS, $F = 10$ THRESHOLD

	(1)
Proportion above 10 in 10±25	0.770
One Sided p-value	0.000
Number of Tests in 10±25	1210
Proportion above 10 in 10±20	0.750
One Sided p-value	0.000
Number of Tests in 10±20	1088
Proportion above 10 in 10±15	0.720
One Sided p-value	0.000
Number of Tests in 10±15	968
Proportion above 10 in 10±10	0.650
One Sided p-value	0.000
Number of Tests in 10±10	775
Proportion above 10 in 10±5	0.550
One Sided p-value	0.0240
Number of Tests in 10±5	373
Proportion above 10 in 10±2.5	0.490
One Sided p-value	0.622
Number of Tests in 10±2.5	165
Proportion above 10 in 10±1	0.450
One Sided p-value	0.825
Number of Tests in 10±1	73

Note: In this table, we present the results of binomial proportion tests where a success is defined as a first stage F-statistic above 10. Reported p-values are the probability of the observed (or greater) proportion given a hypothesized equal probability of being just above and below the threshold. No weights have been applied.

TABLE A26—SIGNIFICANT AT THE 5% LEVEL, ROLE OF EVENT-STUDY GRAPHS FOR DID

	(1)	(2)	(3)	(4)	(5)	(6)
DID with Graph	-0.031	-0.035	-0.028	-0.042	-0.040	0.014
	(0.060)	(0.061)	(0.060)	(0.061)	(0.065)	(0.073)
IV	-0.017	-0.018	-0.003	-0.003	0.009	0.072
	(0.059)	(0.059)	(0.057)	(0.057)	(0.061)	(0.066)
RDD	-0.061	-0.058	-0.051	-0.068	-0.058	-0.005
	(0.066)	(0.068)	(0.066)	(0.066)	(0.069)	(0.075)
RCT	-0.119	-0.115	-0.076	-0.083	-0.082	-0.017
	(0.058)	(0.059)	(0.056)	(0.057)	(0.062)	(0.071)
Top 5		-0.050	-0.006			
		(0.045)	(0.085)			
Year=2018		0.021	0.031	0.025	0.011	0.043
		(0.028)	(0.027)	(0.027)	(0.030)	(0.035)
Experience		-0.002	-0.006	-0.006	-0.006	0.009
		(0.007)	(0.007)	(0.007)	(0.008)	(0.009)
Experience^2		-0.005	0.005	0.006	0.014	-0.028
		(0.018)	(0.018)	(0.019)	(0.020)	(0.025)
Top Institution		0.019	0.026	0.025	-0.002	-0.005
		(0.050)	(0.044)	(0.043)	(0.046)	(0.055)
PhD Top Institution		-0.011	-0.030	-0.022	0.024	0.067
		(0.039)	(0.037)	(0.038)	(0.040)	(0.048)
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	5,202	5,202	5,202	5,202	3,798	2,273
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]
RCT Sig Rate	0.47	0.47	0.47	0.47	0.48	0.49

Note: This table reports marginal effects from probit regressions (Equation (2)). The omitted group is test statistics from DID articles without an event-study graph. The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

TABLE A27—WORKING PAPERS

	(1)	(2)	(3)	(4)
DID	-0.316	-0.366	-0.399	-0.376
	(0.050)	(0.049)	(0.046)	(0.043)
IV	-0.378	-0.430	-0.469	-0.450
	(0.046)	(0.046)	(0.042)	(0.039)
RDD	-0.329	-0.401	-0.434	-0.406
	(0.107)	(0.094)	(0.082)	(0.082)
Top 5		0.003	-0.151	-0.105
		(0.048)	(0.102)	(0.102)
Year=2018		0.192	0.196	0.180
		(0.034)	(0.032)	(0.030)
Experience		0.005	0.003	0.005
		(0.010)	(0.009)	(0.008)
Experience^2		-0.013	-0.007	-0.014
		(0.026)	(0.025)	(0.024)
Top Institution		-0.026	0.003	0.007
		(0.076)	(0.059)	(0.056)
PhD Top Institution		0.027	0.026	0.037
		(0.059)	(0.050)	(0.048)
Reporting Method		Y	Y	Y
Solo Authored		Y	Y	Y
Share Female Authors		Y	Y	Y
Editor		Y	Y	Y
Field FE			Y	
Journal FE				Y
Articles	684	684	684	678

Note: This table reports marginal estimates from a probit model. The dependent variable is a dummy that takes a value one if a published article has a public working paper. The only statistically significant coefficients are DID, IV, RDD and year. Article weights applied.

TABLE A28—WORKING PAPERS

	(1) ALL	(2) DID	(3) IV	(4) RCT	(5) RDD
Published Version	-0.018	-0.049	-0.017	-0.011	-0.030
	(0.016)	(0.051)	(0.040)	(0.037)	(0.033)
Constant	0.549	0.594	0.541	0.505	0.597
	(0.008)	(0.031)	(0.024)	(0.022)	(0.024)
Test Statistics	4,305	852	867	1,155	422
Articles	251	86	86	61	32
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]

Note: This table reports estimates from a linear probability regression with article fixed effects. The dependent variable is a dummy that takes a value one if a given test statistic is significant at the 5% level (i.e. equal to 1.96). The independent variable of interest is a dummy that takes the value of one if a given test statistic is from the published version of an article. The sample is accordingly restricted to estimates from published articles that had an associated working paper. We use the inverse of the number of tests presented in the same article to weight observations.

TABLE A29—SIGNIFICANT AT THE 5% LEVEL

	(1)	(2)	(3)	(4)
DID	0.158	0.149	0.136	0.132
	(0.028)	(0.028)	(0.030)	(0.029)
IV	0.206	0.196	0.185	0.181
	(0.026)	(0.027)	(0.028)	(0.028)
RDD	0.099	0.092	0.083	0.067
	(0.038)	(0.040)	(0.041)	(0.039)
Top 5		0.053	0.039	
		(0.030)	(0.074)	
Year=2018		0.017	0.020	0.021
		(0.022)	(0.022)	(0.022)
Experience		-0.002	-0.004	-0.004
		(0.006)	(0.006)	(0.006)
Experience^2		0.008	0.013	0.014
		(0.019)	(0.019)	(0.019)
Top Institution		-0.059	-0.055	-0.064
		(0.034)	(0.034)	(0.034)
PhD Top Institution		-0.006	-0.014	-0.007
		(0.031)	(0.031)	(0.031)
Reporting Method		Y	Y	Y
Solo Authored		Y	Y	Y
Share Female Authors		Y	Y	Y
Editor		Y	Y	Y
Field FE			Y	
Journal FE				Y
Observations	21,740	21,740	21,740	21,740
RCT Sig Rate	0.37	0.37	0.37	0.37

Note: This table reports marginal effects from probit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table A30—Significant at the 10% Level

	(1)	(2)	(3)	(4)
DID	0.140	0.131	0.127	0.126
	(0.026)	(0.026)	(0.027)	(0.027)
IV	0.201	0.191	0.186	0.183
	(0.024)	(0.025)	(0.026)	(0.026)
RDD	0.074	0.064	0.063	0.054
	(0.034)	(0.035)	(0.036)	(0.035)
Top 5		0.062	0.036	
		(0.028)	(0.070)	
Year=2018		0.016	0.018	0.019
		(0.021)	(0.020)	(0.020)
Experience		-0.001	-0.003	-0.003
		(0.006)	(0.006)	(0.006)
Experience^2		0.005	0.009	0.011
		(0.017)	(0.017)	(0.017)
Top Institution		-0.045	-0.043	-0.048
		(0.033)	(0.032)	(0.032)
PhD Top Institution		-0.030	-0.034	-0.029
		(0.030)	(0.029)	(0.029)
Reporting Method		Y	Y	Y
Solo Authored		Y	Y	Y
Share Female Authors		Y	Y	Y
Editor		Y	Y	Y
Field FE			Y	
Journal FE				Y
Observations	21,740	21,740	21,740	21,740
RCT Sig Rate	0.45	0.45	0.45	0.45

Note: This table reports marginal effects from probit regressions (Equation (2)). 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 article to weight observations.

TABLE A31—SIGNIFICANT AT THE 1% LEVEL

	(1)	(2)	(3)	(4)
DID	0.132	0.128	0.124	0.119
	(0.031)	(0.032)	(0.033)	(0.033)
IV	0.158	0.152	0.146	0.139
	(0.030)	(0.031)	(0.032)	(0.032)
RDD	0.103	0.099	0.095	0.079
	(0.042)	(0.043)	(0.045)	(0.044)
Top 5		0.073	0.055	
		(0.031)	(0.085)	
Year=2018		0.011	0.013	0.018
		(0.024)	(0.024)	(0.024)
Experience		-0.000	-0.002	-0.001
		(0.007)	(0.007)	(0.007)
Experience^2		0.006	0.008	0.006
		(0.022)	(0.022)	(0.022)
Top Institution		-0.067	-0.063	-0.077
		(0.038)	(0.038)	(0.037)
PhD Top Institution		0.008	0.004	0.009
		(0.034)	(0.034)	(0.034)
Reporting Method		Y	Y	Y
Solo Authored		Y	Y	Y
Share Female Authors		Y	Y	Y
Editor		Y	Y	Y
Field FE			Y	
Journal FE				Y
Observations	21,740	21,740	21,740	21,740
RCT Sig Rate	0.25	0.25	0.25	0.25

Note: This table reports marginal effects from probit regressions (Equation (2)). 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 article to weight observations.

Table A32—Significant at the 5% Level, Ambiguous Removed

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.096	0.091	0.056	0.053	0.053	0.029
	(0.034)	(0.033)	(0.032)	(0.033)	(0.038)	(0.048)
IV	0.102	0.097	0.072	0.079	0.092	0.092
	(0.034)	(0.034)	(0.033)	(0.034)	(0.038)	(0.045)
RDD	0.060	0.059	0.026	0.016	0.023	0.006
	(0.047)	(0.048)	(0.046)	(0.047)	(0.049)	(0.055)
Top 5		-0.045	-0.013			
		(0.045)	(0.082)			
Year=2018		0.019	0.028	0.022	0.009	0.043
		(0.028)	(0.027)	(0.028)	(0.030)	(0.035)
Experience		-0.003	-0.007	-0.006	-0.006	0.008
		(0.007)	(0.007)	(0.007)	(0.008)	(0.009)
Experience^2		-0.004	0.007	0.008	0.016	-0.027
		(0.018)	(0.018)	(0.019)	(0.021)	(0.025)
Top Institution		0.014	0.021	0.019	-0.005	-0.005
		(0.050)	(0.045)	(0.044)	(0.047)	(0.055)
PhD Top Institution		-0.012	-0.031	-0.024	0.022	0.068
		(0.040)	(0.038)	(0.038)	(0.040)	(0.049)
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	5,131	5,131	5,131	5,131	3,743	2,241
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]
RCT Sig Rate	0.47	0.47	0.47	0.47	0.48	0.49

Note: This table reports marginal effects from probit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. We use the inverse of the number of tests presented in the same article to weight observations.

TABLE A33—SIGNIFICANT AT THE 5% LEVEL, SINGLE METHOD ARTICLES

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.086	0.080	0.043	0.042	0.033	0.014
	(0.036)	(0.035)	(0.034)	(0.035)	(0.038)	(0.049)
IV	0.110	0.105	0.079	0.089	0.095	0.101
	(0.036)	(0.036)	(0.035)	(0.036)	(0.040)	(0.048)
RDD	0.077	0.073	0.036	0.027	0.026	0.002
	(0.050)	(0.051)	(0.048)	(0.049)	(0.050)	(0.058)
Top 5		-0.050	-0.030			
		(0.049)	(0.086)			
Year=2018		0.019	0.028	0.021	0.008	0.042
		(0.029)	(0.028)	(0.029)	(0.031)	(0.036)
Experience		-0.003	-0.008	-0.006	-0.007	0.010
		(0.007)	(0.007)	(0.007)	(0.008)	(0.009)
Experience^2		-0.004	0.007	0.007	0.016	-0.033
		(0.019)	(0.019)	(0.020)	(0.022)	(0.027)
Top Institution		0.017	0.022	0.019	-0.008	-0.014
		(0.053)	(0.047)	(0.045)	(0.048)	(0.058)
PhD Top Institution		-0.021	-0.039	-0.028	0.019	0.067
		(0.041)	(0.038)	(0.039)	(0.041)	(0.050)
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Field FE			Y			
Journal FE				Y	Y	Y
Observations	4,485	4,485	4,485	4,485	3,292	1,977
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.35]	[1.96±0.20]
RCT Sig Rate	0.46	0.46	0.46	0.46	0.47	0.48

Note: This table reports marginal effects from probit regressions (Equation (2)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. We use the inverse of the number of tests presented in the same article to weight observations.

Table A34—Article and Author Characteristics for Authors Using More than One Method

	(1) DID	(2) IV	(3) RCT	(4) RDD	(5) 2015	(6) 2018	(7) Top 5	(8) Non Top 5	(9) Total
Top 5	0.11	0.19	0.30	0.00	0.12	0.24	1.00	0.00	0.19
	(0.31)	(0.39)	(0.46)	(0.05)	(0.33)	(0.43)	(0.00)	(0.00)	(0.39)
Editor Present	0.56	0.60	0.76	0.52	0.65	0.63	0.72	0.62	0.64
	(0.50)	(0.49)	(0.43)	(0.50)	(0.48)	(0.48)	(0.45)	(0.49)	(0.48)
Solo-Authored	0.28	0.32	0.14	0.57	0.35	0.20	0.25	0.28	0.27
	(0.45)	(0.47)	(0.34)	(0.50)	(0.48)	(0.40)	(0.43)	(0.45)	(0.44)
Average Experience	9.73	10.42	12.91	7.39	9.21	12.17	11.64	10.58	10.78
	(5.11)	(6.44)	(5.31)	(3.01)	(5.54)	(5.34)	(5.93)	(5.54)	(5.63)
Female Authors	0.19	0.26	0.45	0.04	0.22	0.34	0.39	0.26	0.28
	(0.27)	(0.32)	(0.32)	(0.13)	(0.29)	(0.34)	(0.41)	(0.29)	(0.32)
Top Institutions	0.18	0.46	0.49	0.49	0.40	0.40	0.66	0.34	0.40
	(0.31)	(0.42)	(0.35)	(0.43)	(0.42)	(0.37)	(0.36)	(0.38)	(0.39)
Top PhD Institutions	0.37	0.44	0.53	0.55	0.42	0.51	0.56	0.44	0.47
	(0.42)	(0.45)	(0.37)	(0.47)	(0.43)	(0.41)	(0.40)	(0.43)	(0.42)
Test Statistics	1098	901	1444	475	1843	2075	433	3485	3918

Note: Each observation is a test. The Top 5 journals in economics are the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics* and *Review of Economic Studies*. Average experience is the mean of years since PhD for an article's authors. Share of female authors, share of authors affiliated with top institutions, and share of authors who completed a PhD at a top institution.

TABLE A35—ROBUSTNESS CHECK: OMISSION OF ECONOMIC SUB-FIELDS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
DID	0.036	0.045	0.055	0.058	0.069	0.045	0.056	0.033	0.056
	(0.038)	(0.033)	(0.033)	(0.039)	(0.034)	(0.033)	(0.034)	(0.039)	(0.033)
IV	0.057	0.090	0.080	0.094	0.089	0.071	0.086	0.082	0.078
	(0.036)	(0.035)	(0.033)	(0.038)	(0.036)	(0.034)	(0.034)	(0.036)	(0.034)
RDD	0.005	0.025	0.017	0.018	0.021	0.010	0.015	0.012	0.019
	(0.051)	(0.048)	(0.046)	(0.052)	(0.047)	(0.046)	(0.048)	(0.057)	(0.048)
Year=2018	0.033	0.009	0.018	0.048	0.011	0.032	0.029	0.030	0.019
	(0.028)	(0.028)	(0.028)	(0.032)	(0.029)	(0.027)	(0.028)	(0.031)	(0.028)
Experience	-0.016	-0.002	-0.004	-0.004	-0.007	-0.004	-0.005	-0.007	-0.004
	(0.007)	(0.008)	(0.007)	(0.008)	(0.008)	(0.007)	(0.007)	(0.008)	(0.007)
Experience^2	0.032	-0.006	0.004	0.001	0.011	0.005	0.004	0.012	0.004
	(0.020)	(0.020)	(0.019)	(0.021)	(0.024)	(0.019)	(0.019)	(0.021)	(0.019)
Top Institution	0.031	0.053	0.023	-0.031	0.042	0.046	0.030	0.013	0.011
	(0.044)	(0.047)	(0.043)	(0.051)	(0.046)	(0.043)	(0.044)	(0.048)	(0.044)
PhD Top Institution	0.013	-0.021	-0.024	-0.043	-0.022	-0.039	-0.023	-0.033	-0.014
	(0.039)	(0.041)	(0.038)	(0.045)	(0.039)	(0.038)	(0.038)	(0.044)	(0.038)
Reporting Method	Y	Y	Y	Y	Y	Y	Y	Y	Y
Solo Authored	Y	Y	Y	Y	Y	Y	Y	Y	Y
Share Female Authors	Y	Y	Y	Y	Y	Y	Y	Y	Y
Editor	Y	Y	Y	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	4,343	4,411	5,162	3,720	4,566	5,185	5,052	4,303	5,009
Dropped	Top 5	Finance	Macro	Gen. Int.	Dev.	Exp.	Labor	Public	Urban
RCT Sig Rate	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47

Note: This table reports marginal effects from probit regressions (Equation (2)). 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 article to weight observations.