Online Appendix for: Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment

Sebastian Kranz* and Peter Pütz†

Appendix A: Monte-carlo study for comparison with derounding approaches

Instead of omitting coarsely rounded observations, earlier studies such as Brodeur et al. (2016) and Bruns et al. (2019) deround reported coefficients and standard errors by assuming missing digits are drawn from a uniform distribution. Brodeur et. al (2016) randomly draw a single derounded data set. Bruns et al. (2019) reduce noise by repeating this derounding procedure several times. We adapt that method as follows: We draw 100 derounding samples and take the median of the estimated significance share and construct confidence intervals by taking the median of the lower and upper bounds of the 100 confidence intervals.[1]

To compare the different approaches, we perform a Monte-Carlo study with two scenarios. In the first scenario, the simulated true z-statistics are uniformly distributed on the interval $[0, 2 \cdot 1.96]$. In the second scenario, 35% of z-statistics are uniformly distributed on $[0, 1.96]$ and 65% uniformly on $[1.96, 2 \cdot 1.96]$, i.e. in

*Ulm University, Department of Mathematics and Economics, Helmholtzstr. 18, D-89081 Ulm, Germany, sebastian.kranz@uni-ulm.de

†Bielefeld University, Faculty of Business Administration and Economics, Universitätsstr. 25, D-33615 Bielefeld, Germany, peter.puetz@uni-bielefeld.de

[1]This construction is akin to a procedure proposed by of Chernozhukov et al. (2020). For a different application they can establish that the resulting median of the 95% CI bounds has a coverage of at least 90% taking into account the resampling noise. Taking into account the promising results of the Monte-Carlo simulations for the case that the null hypothesis is satisfied, we do not adapt the confidence intervals, however.

Table A1: Results of Monte-Carlo simulations

	Share significant: 50%					Share significant: 65%				
Approach	Bias	95% CI		Cover-age	RMSE	Bias	95% CI		Cover-age	RMSE
No adjustment	0.039	0.515	0.562	10.2%	0.041	0.025	0.653	0.697	38.8%	0.028
Omit $s < 37$	0.000	0.471	0.529	95.0%	0.015	-0.001	0.621	0.676	95.1%	0.014
Derounding assuming uniform distribution of unobserved digits										
Single sample	0.000	0.476	0.524	95.0%	0.012	-0.018	0.609	0.655	66.8%	0.021
Median	0.000	0.476	0.524	97.1%	0.011	-0.018	0.609	0.655	69.0%	0.021

Note: Different approaches to deal with rounding uncertainty are compared with regard to their performance in two different simulation scenarios. The first scenario corresponds to a research field without p-hacking or publication bias, while the second scenario corresponds to a research field with substantial p-hacking or publication bias. 100,000 repetitions are used.

each window around 1.96, we would expect 65% of the tests to be significant. For the numbers of observations and the distribution of significant digits and scaling, we follow closely BCH's pooled data set (see KP for details).

Table A1 shows the average results for 100,000 repetitions for a window with half-width $h = 0.2$ around $z = 1.96$. Not performing rounding adjustment causes a substantial upward bias in the estimated share of significant tests in both scenarios. In contrast, our omission approach seems roughly unbiased and the confidence intervals achieve 95% coverage probability in both scenarios.

The median derounding approach has the lowest root mean squared error (RMSE) in the first scenario and achieves 97.1% coverage probability for the 95% confidence intervals. The rough intuition is that this derounding approach makes the z-statistics more equally distributed, i.e. it causes the sample share of significant tests to be closer to 50%. It thus allows a conservative test for the null hypothesis of a 50% probability of significant tests. The single sample approach of Brodeur et al. (2016) induces more noise that countervails the excess coverage probability.

Unfortunately, both uniform derounding approaches can induce an attenuation bias if the null hypothesis is violated. This bias can be seen in the second scenario, where also the coverage probability of the 95% confidence intervals drops below 70%.

In KP we provide more intuition for those results and also propose derounding

methods that alleviate the attenuation bias. Yet, those methods are more complex and not clearly superior to the simpler omission approach. Thus, we only show results for the omission approach in our main analysis. The more detailed analysis in KP shows that applied to BCH's data set all adjustments for the rounding problem yield very similar results.

Appendix B: Power loss of omission approach

Table A2 provides some insight into the power loss from our rounding adjustment that omits too coarsely rounded observations. On average the widths of the 95% confidence intervals for the share of z-statistics above 1.96 increase by 25.4% in the pooled data.

Table A2: Power loss induced by omission approach

	Average width of 95% CI		Average width
	original sample	adjusted sample	increase
Pooled	0.046	0.059	25.4%
DID	0.085	0.118	36.5%
IV	0.093	0.103	11.3%
RCT	0.084	0.109	27.4%
RDD	0.136	0.183	32.5%

Note: We compute for every window half-width $h \in \{0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$ considered by BCH the width of the 95% confidence interval (using exact binomial test) for the share of z-statistics above 1.96. The table shows the average widths of the confidence intervals (averaged over all h) for the original data set and our adjusted data set where we drop all observations with $s < 37$. The last column shows the average of the percentage increase of the confidence interval width when dropping observations (averaged over all h).

The width of the confidence intervals increase strongest for DID (on average by 36.5%) which is consistent with the fact that the DID sample has the largest share of observations with $z = 2$ and thus seems to be most affected by coarse rounding. For the DID sample, the increases in the widths of the confidence intervals range from 60.1% for $h = 0.05$ to 26.8% for $h = 0.3$.

Note that initial sample sizes differed substantially between methods with RDD having less than half of the number of observations than RCT. Even before our adjustment the confidence intervals for the RDD sample were wider than those

of any other method before or after adjustment.

While our adjustment increases the width of the confidence intervals, recall from the Monte-Carlo simulation in Appendix A that in both scenarios the adjustment leads to a more precise estimate in terms of RMSE than the biased estimator without adjustment.

Appendix C: Proof of Lemma 1

Lemma 1. *Assume reported standard error σ and coefficient μ are rounded to the same number of decimal places. The true and reported z-statistics $\tilde{z}$ and z are guaranteed not to lie on opposite sides of an arbitrary threshold τ if the significand s of the standard error satisfies*

$$s \geq \frac{1+\tau}{2|z-\tau|}.$$

Proof. The smallest and largest possible values of $\tilde{z}$ are given by

$$\tilde{z}_{min} = \frac{zs - 0.5}{s + 0.5} \text{ and } \tilde{z}_{max} = \frac{zs + 0.5}{s - 0.5}.$$

If $z \geq \tau$, we need $\tilde{z}_{\min} \geq \tau$ to guarantee $\tilde{z} \geq \tau$, which can be rearranged to

$$s \geq \frac{1+\tau}{2|z-\tau|}.$$

In a similar spirit if $z \leq \tau$, the relevant condition is $\tilde{z}_{max} \leq \tau$, which also can be rearranged to the same condition. □

Appendix D: Randomization tests and caliper tests at 10% and 1% significance thresholds

Tables A3 and A4 show our replication results of the randomization tests using the adjusted data set at the 10% and 1% significance threshold, respectively. Following BCH, we report one-sided p-values at the 10% threshold ($z = 1.65$). At the 1% threshold ($z = 2.58$) we report two-sided p-values because we find in many instances that the share of z-statistics above the threshold is significantly

below 50%.

Table A3: Randomization tests, 10% significance threshold $z = 1.65$

	(1) ALL	(2) DID	(3) IV	(4) RDD	(5) RCT
Window half-width 0.05					
Proportion significant	0.556	0.512	0.62	0.5	0.552
(p-value)	(0.025)	(0.456)	(0.010)	(0.560)	(0.179)
Observations	320	80	100	44	96
Window half-width 0.075					
Proportion significant	0.54	0.562	0.573	0.456	0.531
(p-value)	(0.040)	(0.101)	(0.047)	(0.802)	(0.240)
Observations	494	121	143	68	162
Window half-width 0.1					
Proportion significant	0.553	0.569	0.591	0.466	0.543
(p-value)	(0.004)	(0.057)	(0.007)	(0.772)	(0.112)
Observations	644	144	193	88	219
Window half-width 0.2					
Proportion significant	0.555	0.585	0.584	0.503	0.53
(p-value)	(0.000)	(0.001)	(0.001)	(0.500)	(0.102)
Observations	1,372	325	385	177	485
Window half-width 0.3					
Proportion significant	0.561	0.61	0.575	0.518	0.534
(p-value)	(0.000)	(0.000)	(0.000)	(0.293)	(0.039)
Observations	2,036	467	581	274	714
Window half-width 0.4					
Proportion significant	0.566	0.617	0.602	0.501	0.525
(p-value)	(0.000)	(0.000)	(0.000)	(0.500)	(0.065)
Observations	2,767	630	816	353	968
Window half-width 0.5					
Proportion significant	0.561	0.618	0.603	0.493	0.514
(p-value)	(0.000)	(0.000)	(0.000)	(0.630)	(0.172)
Observations	3,462	781	1,023	446	1,212

Note: Replicates Table A6 in BCH. We present for several windows centered around z=1.65 the proportion of significant observations and test if it is statistically greater than 0.5.

Table A4: Randomization tests, 1% significance threshold $z = 2.58$

	(1) ALL	(2) DID	(3) IV	(4) RDD	(5) RCT
Window half-width 0.05					
Proportion significant	0.462	0.351	0.514	0.406	0.527
(p-value)	(0.217)	(0.012)	(0.847)	(0.377)	(0.728)
Observations	290	77	107	32	74
Window half-width 0.075					
Proportion significant	0.474	0.404	0.51	0.438	0.514
(p-value)	(0.306)	(0.049)	(0.871)	(0.471)	(0.847)
Observations	420	114	151	48	107
Window half-width 0.1					
Proportion significant	0.464	0.412	0.487	0.422	0.503
(p-value)	(0.093)	(0.035)	(0.776)	(0.260)	(1.000)
Observations	567	153	197	64	153
Window half-width 0.2					
Proportion significant	0.432	0.411	0.44	0.387	0.463
(p-value)	(0.000)	(0.004)	(0.023)	(0.010)	(0.225)
Observations	1,087	275	375	137	300
Window half-width 0.3					
Proportion significant	0.417	0.41	0.416	0.396	0.432
(p-value)	(0.000)	(0.000)	(0.000)	(0.005)	(0.005)
Observations	1,598	412	548	187	451
Window half-width 0.4					
Proportion significant	0.408	0.422	0.4	0.399	0.409
(p-value)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Observations	2,130	555	723	258	594
Window half-width 0.5					
Proportion significant	0.388	0.416	0.386	0.375	0.37
(p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	2,700	695	924	333	748

Note: Replicates Table A7 in BCH. We present for several windows centered around z=2.58 the proportion of significant observations and test if it is statistically equal to 0.5. Thus, in contrast to Table A7 in BCH, two-sided p-values are shown.

Tables A5 and A6 show our replication results of the caliper tests using the adjusted data set at the 10% and 1% significance threshold, respectively.

Table A5: Caliper tests, 10% significance threshold $z = 1.65$

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.100	0.113	0.108	0.111	0.103	0.090
	(0.044)	(0.040)	(0.041)	(0.041)	(0.047)	(0.060)
IV	0.116	0.126	0.101	0.111	0.066	0.039
	(0.036)	(0.038)	(0.040)	(0.041)	(0.043)	(0.053)
RDD	0.015	0.009	0.004	0.005	-0.024	-0.069
	(0.042)	(0.041)	(0.044)	(0.045)	(0.056)	(0.062)
Top 5		0.060	-0.047			
		(0.049)	(0.102)			
Year=2018		0.003	0.010	0.020	0.024	0.078
		(0.035)	(0.034)	(0.035)	(0.037)	(0.042)
Experience		-0.002	-0.006	-0.007	-0.016	-0.011
		(0.009)	(0.009)	(0.008)	(0.008)	(0.009)
Experience squared		-0.003	0.008	0.009	0.032	0.020
		(0.027)	(0.026)	(0.024)	(0.023)	(0.025)
Top institution		-0.057	-0.044	-0.039	-0.006	-0.001
		(0.049)	(0.049)	(0.048)	(0.052)	(0.063)
Top PhD institution		-0.016	-0.032	-0.045	-0.059	-0.147
		(0.045)	(0.045)	(0.043)	(0.049)	(0.061)
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	3,470	3,470	3,470	3,469	2,377	1,363
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	.52	.52	.52	.52	.53	.54

Note: Replicates Table A16 in BCH. The shown coefficients are marginal effects at the means. For dummy variables we measure the effect of a change from 0 to 1. Standard errors in parentheses are clustered at article level. Observations are weighted by the inverse of the number of tests conducted in the same article.

Table A6: Caliper tests, 1% significance threshold $z = 2.58$

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.050	0.040	0.021	0.010	0.027	-0.062
	(0.049)	(0.045)	(0.048)	(0.048)	(0.054)	(0.063)
IV	0.019	0.009	-0.017	-0.039	-0.023	-0.016
	(0.042)	(0.039)	(0.040)	(0.042)	(0.046)	(0.060)
RDD	-0.082	-0.083	-0.094	-0.092	-0.071	-0.122
	(0.068)	(0.062)	(0.062)	(0.065)	(0.069)	(0.096)
Top 5		0.029	-0.101			
		(0.056)	(0.101)			
Year=2018		0.010	0.008	0.022	0.018	0.013
		(0.041)	(0.041)	(0.041)	(0.043)	(0.047)
Experience		0.018	0.016	0.013	0.013	0.004
		(0.009)	(0.009)	(0.009)	(0.011)	(0.013)
Experience squared		-0.049	-0.044	-0.037	-0.038	-0.020
		(0.030)	(0.029)	(0.028)	(0.039)	(0.047)
Top institution		0.003	0.006	0.006	-0.008	0.064
		(0.062)	(0.063)	(0.062)	(0.066)	(0.073)
Top PhD institution		-0.051	-0.053	-0.038	0.030	0.010
		(0.053)	(0.054)	(0.056)	(0.063)	(0.069)
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	2,689	2,689	2,689	2,689	1,850	1,079
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	.37	.37	.37	.37	.42	.46

Note: Replicates Table A18 in BCH. The shown coefficients are marginal effects at the means. For dummy variables we measure the effect of a change from 0 to 1. Standard errors in parentheses are clustered at article level. Observations are weighted by the inverse of the number of tests conducted in the same article.

Appendix E: Additional figures

Figure A1 replicates Figure 5 in BCH using our adjusted data set. It presents the distribution of first stage F-statistics (top panel) and associated second stage z-statistics (lower panels) from IV analyses, the latter split by relatively weak and relatively strong instruments. The results are similar to BCH, only the distribution of test statistics for IVs with relatively low F-statistics (bottom left panel) exhibits a slightly more pronounced peak just above 1.96. Put differently, IV studies with comparatively weak instruments have an even higher proportion of

z-statistics in the second stage that are around or above conventional significance thresholds.

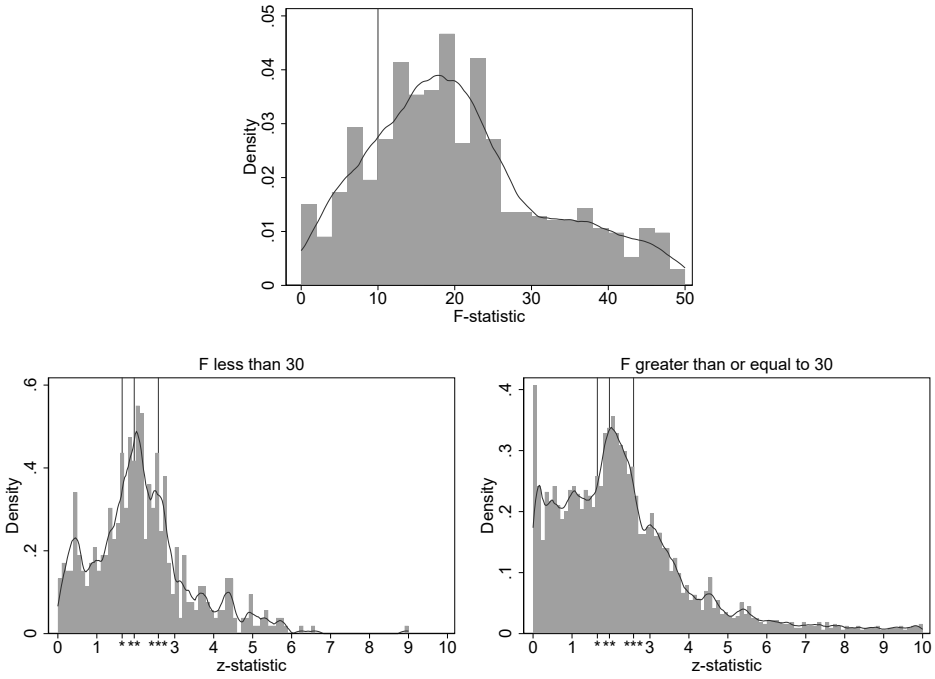

Figure A1: Instrumental variables: first stage F-statistics and associated second stage z-statistics

Note: Replicates Figure 5 in BCH. The upper panel displays the distribution of first stage F-statistics of instrumental variables for F $\in$ [0, 50]. Histogram bins have a width of 2. A reference line is provided at the conventional threshold of 10 for "weak" instruments. The bottom left panel displays the distribution of second stage z-statistics for IVs with a relatively low first stage F-statistic (below 30), while the bottom right panel displays the distribution of second stage z-statistics for IVs with a F-statistic above 30. A total of 1,082 statistics are used in this analysis. The bottom left panel contains 531 tests, while the bottom right contains 551 tests.

Figure A2 compares the test statistics in working papers and the respective journal versions.[2] The figure looks very similar to Figure 6 in BCH with a slightly

[2]The working paper data does not include information on trailing zeros in reported standard errors. To heuristically recover trailing zeros when computing the significand s of the standard error, we exploit the convention that the coefficient and associated standard error are usually reported with the same number of decimal places. If the reported coefficient has more decimal digits than the reported standard errors, zeros are appended to the standard errors until the number of decimals of the associated coefficient is reached. For Figure A2 we apply this heuristic for both the working papers and published articles and then drop in both data sets about 31% of tests by applying our omission approach. To validate the procedure, we tested it on the

more pronounced peak just above 1.96 for the working paper statistics when applying the omission approach.

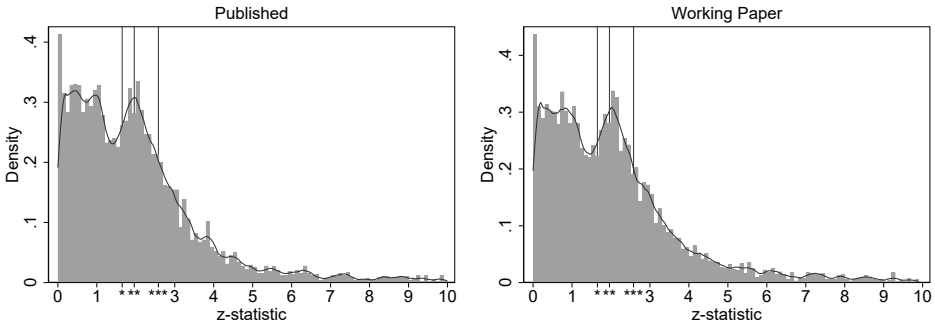

Figure A2: Instrumental variables: first stage F-statistics and associated second stage z-statistics

Note: Replicates Figure 6 in BCH. The figure displays distributions of test statistics for $z \in [0, 10]$. The left panel restricts the sample to journal articles for which working papers could be found, while the right panel contains the z-statistics from the respective working papers. The vertical lines indicate the the critical z-statistics at the 10%, 5% and 1% significance levels. The histograms have bin size 0.1. The black lines are density estimates based on a Epanechnikov kernel with bandwidth 0.1.

Figure A3 shows a variant of Figure 6 in our comment using kernel density estimators that correct the negative bias at $z = 0$.

complete article data set of BCH in which trailing zeros are specified. Our heuristic led in 96% of cases to the same decision to either keep or drop an observation.

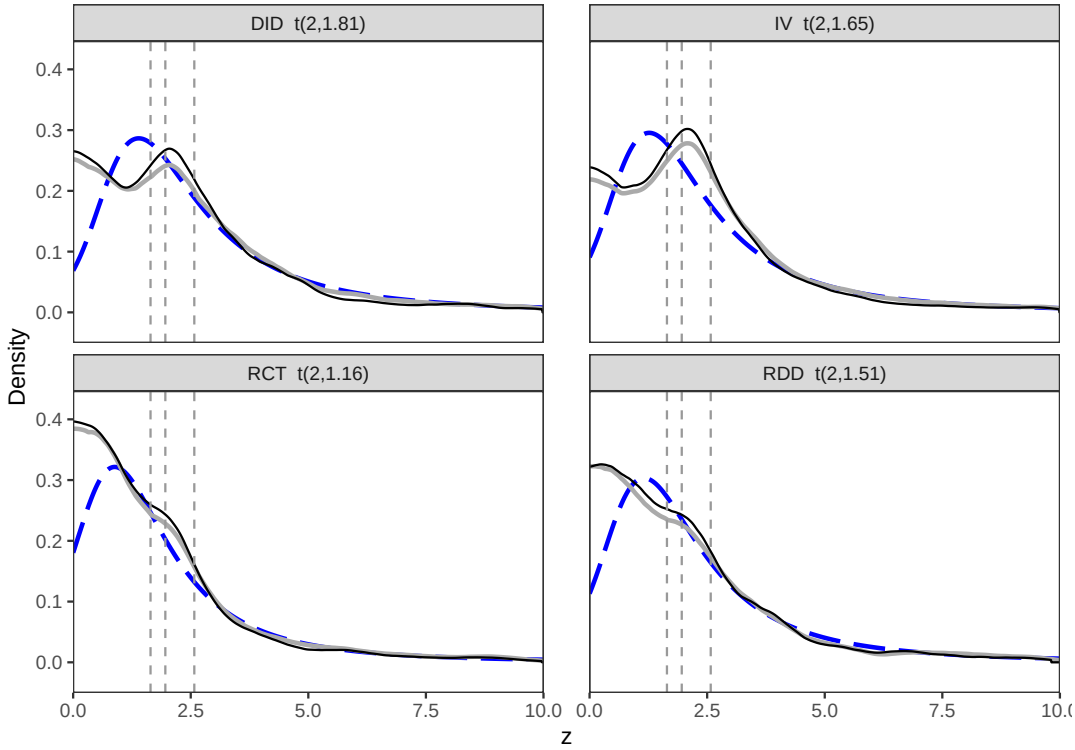

Figure A3: Excess test statistic plots (bias corrected)

Note: This is a version of Figure 6 in our comment (Figure 4 in BCH) that corrects for the downward bias of the kernel density estimator at $z = 0$. We adjust the kernel density estimates by imposing the assumption that the distribution is symmetric for positive and negative z-statistics. This kernel density estimator matches the high empirical density of z-statistics close to zero that can also be observed in the histograms in Figure 4.

Appendix F: Proposition 1 and discussion of alternative calibration based on conditional tail probabilities

BCH calibrate the degrees of freedom and non-centrality parameters of t-distributions by matching the probability mass in the tails ($z > 5$) with those of the empirical distributions, assuming the empirical distributions to be free of p-hacking and publication bias for $z > 5$. The following Proposition shows that BCH's calibration approach is not able to recover the true distribution of z-statistics absent publication bias, even if the correct functional form is assumed.

Proposition 1. *Let $F^*(z)$ be a distribution function of absolute z-statistics absent publication bias. Let $p(z) \in [0,1]$ be the publication probability of a test with z-statistic z. Let $F(z)$ denote the resulting distribution function of observed z-statistics. We assume publication bias is present, i.e. for some $z \geq 0$, we have $F(z) < F^*(z)$. Assume there is a threshold $\bar{z} > 0$ such that $p(z) = 1$ for all $z \geq \bar{z}$. Then for every $z \geq \bar{z}$ we find*

$$1 - F(z) = \mu \left(1 - F^*(z)\right)$$

with $\mu > 1$. This means that the tails of the distribution of observed z-statistics starting at a z-statistic above which no publication bias takes place have a higher probability mass than the corresponding tails in the latent distribution.

Proof. Let $\tilde{F}(z) = \int_0^z p(z) dF^*(z)$. Let $M = \lim_{z\to\infty} \tilde{F}(z)$. Note that M is strictly below 1. Let $\mu = 1/M$. The distribution function of observed z statistics is given by $F(z) = \mu\tilde{F}(z)$. For any pair $z_1 \geq \bar{z}$ and $z_2 \geq z_1$ we have $F^*(z_2) - F^*(z_1) = \tilde{F}(z_2) - \tilde{F}(z_1)$ since $p(z) = 1$ for all $z \geq \bar{z}$. This implies $F(z_2) - F(z_1) = \mu(F^*(z_2) - F^*(z_1))$. The proposition follows from setting $z_1 = z$ and taking the limit $z_2 \to \infty$. □

Technically, under BCH's assumptions the parameters of the latent distribution can be identified and estimated via maximum likelihood from the conditional tail distribution conditioning on $z \geq 5$. Let F_θ describe the cumulative distribution function and f_θ the density function of a non-central t-distribution characterized by a parameter vector θ comprising the degrees of freedom (df) and the non-centrality parameter (ncp). In this alternative approach, we choose the parameter vector θ that maximizes the log likelihood function based on the conditional tail densities

$$l(\theta|\tilde{Z}) = \sum_{z_i \in \tilde{Z}} \log \frac{f_\theta(z_i)}{1 - F_\theta(5)},$$

where $\tilde{Z}$ contains the observed z-statistics in the tail. The practical problem of this approach is that the conditional tail densities can look very similar for very different combinations of the df and ncp parameters. This makes it very hard to get sufficiently precise estimates. To illustrate the problem, we perform that maximum likelihood estimation for different subsamples. We distinguish by the

causal identification strategy and either use all z-statistics in the tail or omit very large z-statistics above the thresholds 1000, 100 or 10.[3] The results are shown in Table A7.

Table A7: Estimated parameters of non-central t-distributions using conditional tail probabilities

	unadjusted sample				adjusted sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	DID	IV	RCT	RDD	DID	IV	RCT	RDD
Sample: all $z \geq 5$								
df	1.4	1.8	1.6	1.2	1.5	2	1.4	2.1
ncp	-5.6	1.6	-4.5	-5.3	2.1	2.8	-5.7	1.7
Observations	902	706	713	380	386	386	376	157
Sample: $5 \leq z \leq 1000$								
df	1.4	1.9	1.6	1.3	1.9	2	1.5	2.1
ncp	-4.7	2.3	-4.3	-0.8	4.2	2.8	-4.4	1.7
Observations	901	705	710	377	385	386	373	157
Sample: $5 \leq z \leq 100$								
df	1.6	2	1.7	1.5	1.9	2.2	1.5	3.2
ncp	1.7	2.8	-0.1	3.2	4.2	3.5	-4.7	4.9
Observations	887	703	707	370	385	384	371	155
Sample: $5 \leq z \leq 10$								
df	12.1	11.6	13.2	15.8	12.3	10.5	15.4	17.3
ncp	5.8	5.7	5.8	6.3	5.9	5.5	6	6.1
Observations	573	492	493	229	246	269	254	119

Note: Columns (1)-(4) show results for the original sample without rounding adjustment and columns (5)-(8) use our sample with adjustment for coarse rounding. The log likelihood function based on the conditional tail densities has been maximized using the Nelder-Mead method.

We see that the estimated values for the non-centrality parameter (ncp) vary hugely depending on which very large z-statistics are included. For example, for

[3]Excluding large outliers is a common practice. For example, BCH only consider z-statistics below 10 when deriving the latent distribution using the approach of Andrews and Kasy (2019).

DID without rounding adjustment (column 1) our estimate for the ncp ranges from a negative -5.6 if all $z \geq 5$ are included to a positive 5.8 if only all $z \in [5, 10]$ are considered. The corresponding latent distributions are very different. Moreover, comparing the first and second sample, we see that just a single additional observation changes the estimate of the ncp parameter by a substantial amount of 0.9 units; in column 5 a single observation changes the ncp estimate even by 2.1 units. Also for the other identification strategies estimates vary largely across the different samples. Likewise, our rounding adjustment sometimes strongly changes the estimated parameters. This exemplifies the practicial difficulties to precisely calibrate the coefficients of the tail distributions using this alternative approach based on conditional tail probabilities.

Appendix G: Results of specification test for Andrews and Kasy (2019) approach

Table A8: Specification test for Andrews and Kasy (2019) approach

	DID	IV	RCT	RDD
Complete sample:				
Panel A	0.92	0.90	0.92	0.92
	[0.92, 0.93]	[0.88, 0.91]	[0.90, 0.94]	[0.91, 0.93]
Panel B	0.92	0.89	0.92	0.92
	[0.91, 0.93]	[0.88, 0.91]	[0.90, 0.94]	[0.91, 0.93]
Sample adjusted for coarse rounding:				
Panel A	0.92	0.90	0.92	0.91
	[0.91, 0.93]	[0.88, 0.91]	[0.89, 0.94]	[0.89, 0.92]
Panel B	0.92	0.89	0.92	0.91
	[0.90, 0.93]	[0.88, 0.91]	[0.89, 0.94]	[0.89, 0.92]

Note: The table shows the inverse probability weighted correlations between $\log \mu$ and $\log \sigma$ as explained in Section 3E. We compute the correlations for each method and for the two specifications of the Andrew and Kasy approach corresponding to Panel A and B in BCH's Table 5. In brackets below the correlations are the bootstrapped 95% confidence intervals for the correlation. For each of the 500 bootstrap samples, we re-apply Andrews and Kasy's (2019) procedure to estimate publication probabilities before computing the inverse probability weighted correlations.

Additional References

- **Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val.** 2020. "Generic machine learning inference on heterogenous treatment effects in randomized experiments, with an application to immunization in India", National Bureau of Economic Research Working Paper 24678.