Rounding and other pitfalls in meta-studies on p-hacking and publication bias: A comment on Brodeur et al. (2020)*

Sebastian Kranz† and Peter Pütz‡

November 2021

Abstract

Brodeur et al. (2020) study hypothesis tests from economic articles and find evidence for p-hacking and publication bias, in particular for IV and DID studies. We adjust for rounding errors in reported estimates and standard errors using different approaches. Statistical evidence for p-hacking from randomization tests and caliper tests at the 5% significance threshold robustly vanishes for DID studies but remains for IV studies. In addition, BCH derive latent distribution of z-statistics absent publication bias using two different approaches. We show that neither approach is applicable due to reasons unrelated to rounding.

Abel Brodeur, Nikolai Cook, and Anthony Heyes (2020), henceforth BCH, have collected a huge, very insightful data set of hypothesis tests from 25 economic journals. They compare articles that employ different empirical strategies to estimate causal effects and find evidence for p-hacking or publication bias overall and in particular for

*This working paper extends two comments that we independently submitted to the *American Economic Review* end 2020 and beginning 2021 and were invited for a merged re-submission. A shortened version of this working paper will be resubmitted to the *American Economic Review*. We thank the editor Isaiah Andrews and two anonymous referees for very helpful comments and suggestions. We also thank the original authors, Abel Brodeur, Nikolai Cook, and Anthony Heyes, for providing us with an updated data set that contains information on trailing zeros and for collecting and sharing information on reported significance levels for a subsample of tests from DID studies. We also thank Maike Hohberg for helpful comments.

†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

results relying on instrumental variables (IV) and to a smaller extent for difference-in-differences estimates (DID).

This paper first shows in Section 1 that a crucial continuity assumption of the underlying z-statistic is violated in the collected data set because of rounding errors in the reported coefficients and standard errors that cause bunching of computed z-statistics. Particularly problematic is that many reported z-statistics bunch at exactly $z = 2$. In Section 2, we compare and partly newly develop different approaches to adjust for rounding errors. All yield similar results on BCH's data set as Section 3 shows: Evidence for p-hacking in the data set substantially weakens. Replicating BCH's randomization and caliper tests at the 5% significance threshold, evidence remains for IV but not for DID and the other methods.

In Section 4, we note issues unrelated to rounding. In their analysis of excess test statistics, BCH use two different approaches to recover the latent distribution of z-statistics absent publication bias and p-hacking. The first approach uses a calibration that relies on the assumption that p-hacking and publication bias don't affect the probability mass in the tails of observed z-statistics (for $z \geq 5$). Yet, we prove that this approach generally fails to recover the true latent distribution. Even if publication bias only reduces publication probabilities for small z, it will affect the probability mass in the tails, as the total probability mass has to integrate to one. We also note that the supposed latent distributions cannot explain an excess share of observed z-statistics close to zero. The second approach to recover a latent distribution is based on Andrews and Kasy (2019). It requires that in the latent distribution coefficients and standard errors are independently distributed from each other. We test this assumption and show that it is strongly violated in BCH's data set.

A feature of BHC's analysis that robustly remains is that the density of z-statistics shows a hump loosely around $z = 2$. In our concluding remarks, we discuss a possible alternative source for this hump and the implications for randomization tests also at the 10% and 1% significance thresholds. Finally, we directly compare the distribution of z-statistics of the different subsamples to that of randomized controlled trials (RCT). The results suggest that insignificant results for DID, IV and to smaller extent for regression discontinuity designs (RDD) face a relatively tougher publication hurdle than for RCT. The code to replicate our analyses is made available on the following GitHub repository: `https://github.com/peterpuetz2020/replication_repo_methods_matter_comment`.

1 The rounding problem

BCH have collected data for more than 21,000 hypothesis tests. For most tests (90.2%) the reported coefficient μ and its standard error σ are collected and the (absolute) z-statistic $z = abs(\mu)/\sigma$ is computed from these values. For the remaining tests, the z-statistic is derived from a reported t-statistic (5.0%), p-value (4.7%) or CI (0.1%).[1]

BCH's main statistical analyses focus on the 5% significance threshold. For their randomization tests, they assume that absent p-hacking or publication bias, the distribution of z-statistics would be continuous and differentiable so that in a sufficiently small window $[1.96 - h, 1.96 + h]$ around the 5% significance threshold, there should be roughly equally many significant results with $z \geq 1.96$ as insignificant results with $z < 1.96$.[2]

BCH compare the shares of significant and insignificant tests for a grid of window half-widths $h \in \{0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$. For the smallest considered window significantly and substantially more significant z-statistics can be found for all subsamples differentiated by strategy for estimating causal effects. For the largest windows significantly more z-statistics above 1.96 are observed only for DID and IV.

The left panel of Figure 1 shows the corresponding shares of z-statistics above 1.96 for all window half-widths on a fine grid between 0.01 and 0.5 using the pooled data.

[1] One of our initial comments also detected that BCH's conversion from p-values into z-statistics wrongly assumed that all p-values correspond to one-sided tests. Brodeur et al. corrected this problem and also detected some smaller typos when converting the raw data. They kindly provided us with a corrected version of the data set and also made it publicly available. We use that updated data set for all our analyses in this comment.

[2] Ideally, one would use observation-specific thresholds that refer to the t distribution and take into account the degrees of freedom instead of using the normal approximation. Pütz and Bruns (2020) show that low degrees of freedom are regularly encountered in top economics. One important reason is the frequent use of clustered standard errors and thus the number of clusters as base for the degrees of freedom. In particular due to clustered standard errors, it is often not possible to collect the correct degrees of freedom from the information given in regression tables.

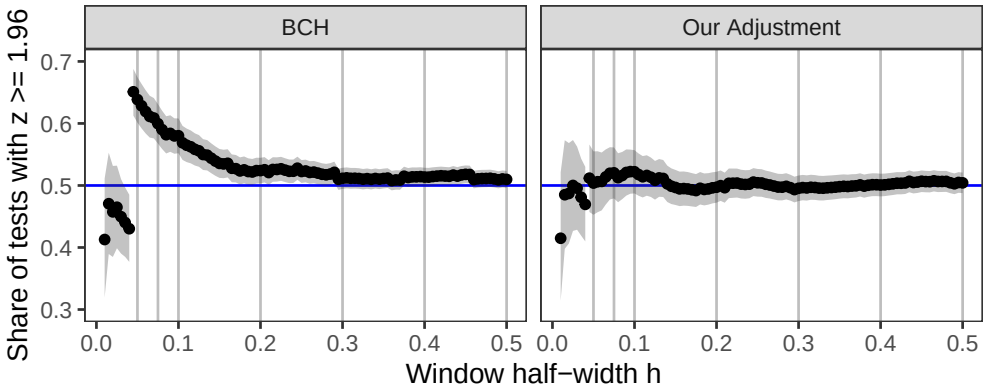

Figure 1: Share of significant results for different window half-widths (pooled data)

Notes: The left panel corresponds to the case of no adjustment for rounding errors as in BCH. The right panel shows results with our adjustment where we omit all observations whose reported standard error has a significand below 37. The shaded areas indicate 95% confidence intervals computed from a t-distribution. The gray vertical lines indicate the window half-widths that BCH studied.

For window half-widths below 0.04, less than 50% of z-statistics are above the threshold of 1.96. But there is a massive, discontinuous increase of significant z-statistics once the window half-width exceeds 0.04. This jump has not been discussed by BCH whose smallest considered window half-width of $h = 0.05$ is already on the right-hand side of this discontinuity.

The discontinuity occurs because the data set contains 260 z-statistics with a value of exactly 2. All these observations are counted as significant and thus cause the jump in the share of significant tests once $1.96 + h$ reaches 2. These 260 observations constitute 37.9% of the total observations in the smallest window analyzed by BCH. The left panel of Figure 2 shows the number of observations included for each window half-width and verifies the substantial jump at $h = 0.04$. The right panels of Figure 1 and Figure 2 present adjusted results and are discussed in the next section.

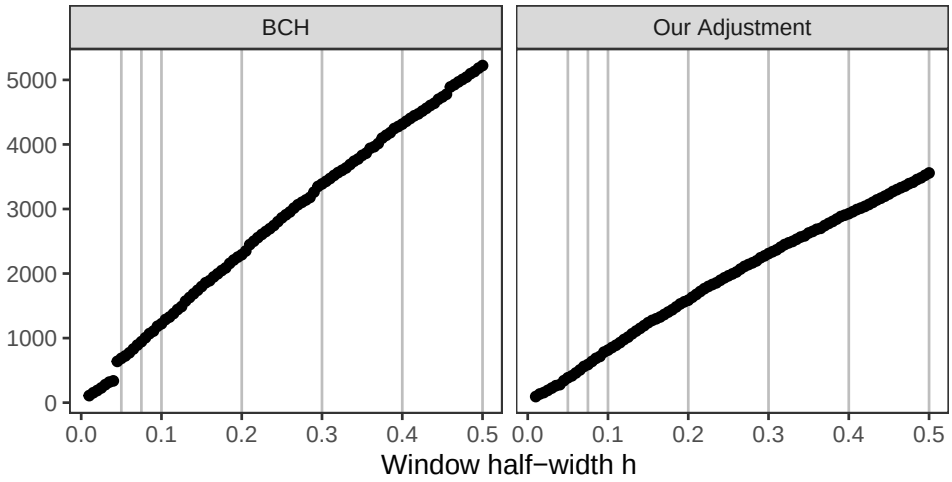

Figure 2: Number of included observations for different window half-widths

Notes: The left panel corresponds to the case of no adjustment for rounding errors as in BCH. The right panel shows results with our adjustment where we omit all observations whose reported standard error has a significand below 37. The gray vertical lines indicate the window half-widths that BCH studied.

Table 1 reveals that rounding errors are likely to be the most important reason for the bunching of z-statistics at exactly $z = 2$. 68.6% of observations with a z-statistic of exactly 2 have just a single significant digit for the standard error and 97.7% have at most two significant digits. For the remaining observations these shares are just 17.2% and 59.8%, respectively.

Table 1: Number of significant digits

Subset of tests	n	Share of standard errors with... 1 significant digit	1 or 2 significant digits
z is exactly 2	260	68.6%	97.7%
other observations	21,480	17.2%	59.8%

If the estimates and standard errors had been reported with more significant digits, the computed z-statistic could well have been smaller than 1.96. For example, assume the reported standard error is $\sigma = 0.02$. Then this observation has a computed z-statistic of exactly $z = 2$ if the reported estimate is also rounded to one significant digit and given by $\mu = 0.04$. If one computed the z-statistic with the original non-rounded

values, it may range from an insignificant lower bound of $z = 1.4$ (i.e. 0.035/0.025) to a highly significant upper bound of $z = 3$ (i.e. 0.045/0.015).

BCH consider all observations with a computed z-statistic above 1.96 as significant at the 5% level, independent of the number of significant digits in the reported estimates and standard errors. But there is no guarantee that in the presence of rounding errors all observations with computed $z \geq 1.96$ were originally reported as significant at the 5% level.[3]

2 Approaches to adjust for the rounding problem

2.1 Omitting observations that are too coarsely rounded

Our first approach for adjusting for rounding errors is derived from a simple idea: just omit all observations that were reported with too few significant digits. For a conservative approach, we would recommend to keep only observations which report estimates and standard errors with at least three significant digits. However, omitting observations reduces power and thus makes it harder to find evidence for p-hacking. In our reanalysis of BCH's results, we will follow a less conservative approach and omit all observations whose standard error has a significand s below 37. The significand consists of the significant digit(s) written as an integer, e.g. for $\sigma = 0.012$ the significand is $s = 12$.[4] In the following, we motivate and illustrate this criterion.

Crucial assumption

A crucial assumption is that the omission criterion is not systematically related to the unobserved distribution of true z-statistics absent rounding errors. E.g. if coarse rounding was more (less) likely if the unobserved true z is above the significance threshold, the adjustment could induce a selection bias causing a systematical under-estimation (over-estimation) of the share of significant test statistics in the randomization tests.

The assumption cannot be directly tested as the unrounded, true z-statistics are unobserved. However, it suggests that for the reported z-statistics the distribution

[3]Replying to our initial comments, Brodeur et al. looked at the DID articles with the largest number of tests with $z = 2$ and collected the reported significance stars. From the collected 43 tests 5 had no stars, 14 had one star, 21 had two stars and 3 had three stars.

[4]Brodeur et al. kindly added to the updated data set new columns where coefficients and standard errors are given as strings that also indicate trailing zeros. We only omit (or perform derounding) among the roughly 90% of observations whose reported z-statistic is computed from reported μ and σ.

Table 2: Omitted observations

Subset	Original observations	Share omitted
$z = 2$	260	87.3%
Other z with at least 10 observations	4,212	81.1%
z with fewer than 10 observations	17,268	26.6%
All observations	21,740	37.9%

of the selected sample should look similar to that of the full sample, except for the bunching points. Figure 5 (further below) shows that indeed the kernel density estimates of the two samples look very similar. In a similar spirit, the assumption suggests that reported z-statistics and reported standard errors should not exhibit a strong correlation: indeed we find a correlation of only -0.0002 with 95% confidence interval $[-0.0138, 0.0134]$. Also the empirical correlation between a dummy indicating that an observation is omitted and the reported z-statistic is only -0.005 with 95% confidence interval $[-0.019, 0.008]$.

One concern could be that coarse rounding might go hand in hand with p-hacking in order to report higher z-statistics. This would suggest that we observe more coarse rounding for z-statistics that are close to the 5% threshold where p-hacking concerns may be most relevant. However, Figure 5 shows that in that range the selected subsample without coarse rounding even has a slightly larger density than the complete sample. Also recall Footnote 3, which shows that almost half of a selection of DID tests with $z = 2$ indeed reported significance level above 5%, i.e. for these tests coarse rounding was not used to wrongfully suggest a 5% significance level.

The omission threshold

On average, a smaller significand s corresponds to a more severe rounding problem for the reported z-statistic. Table 2 shows that our adjustment that omits all observations with $s < 37$ omits 37.9% of all observations, but 87.3% of observations with $z = 2$ and roughly 81.1% of other observations that have a z-statistic on which at least 10 observations were bunched on.

Correspondingly, the right-hand panel of Figure 2 shows that with our sample selection, the number of included observations increases smoothly with growing window half-width h without any visible jumps.

To motivate why we picked the significand $s = 37$ as omission threshold, we continue

with

Lemma 1. *Consider an observation where the reported coefficient μ and standard error σ are rounded to the same decimal place. Let $z = \mu/\sigma$ be the reported z-statistic and $\tilde{z}$ be the unobserved true z-statistic corresponding to the non-rounded values. Then $\tilde{z}$ and z are guaranteed not to lie on opposite sides of an arbitrary threshold τ if the significand s of σ satisfies*

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

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 (1). In a similar spirit if $z \leq \tau$ the relevant condition is $\tilde{z}_{max} \leq \tau$, which also can be rearranged to (1). □

We say an observation is *misclassified* if its reported z-statistic z and true z-statistic $\tilde{z}$ lie on opposite sides of the significance threshold $\tau = 1.96$. Lemma 1 implies that the smallest omission threshold guaranteeing that no observation with $z = 2$ is misclassified is $s = 37$. In total, there are 160 distinct z-statistics with at least 10 observations, but among them $z = 2$ is closest to the significance threshold $\tau = 1.96$. Note from Lemma 1 that the minimally required significand decreases in the distance $|z - \tau|$. Hence, our omission threshold of $s = 37$ implies for all of these 160 bunching values of z that no remaining observation is misclassified.

Besides misclassification, rounding errors can also cause an observation to be wrongly included in a window $[1.96 - h, 1.96 + h]$ if z is inside the window but $\tilde{z}$ outside, or wrongly excluded the other way round.[5] Lemma 1 implies that an omission threshold $s = 37$ guarantees that no remaining observation with $z = 2$ is wrongly included or excluded from windows with half-width $h \geq 0.08$. Thus, only for the two smallest window half-widths considered by BCH, we cannot rule out that after our adjustment some observations with $z = 2$ are still wrongly included in that window. Of course, our adjustment cannot guarantee a zero risk of misclassification, wrong inclusion or exclusion of observations for all reported values of z.

[5]Note that collecting data (if available) on the reported significance levels for each observation could be another way to solve the misclassification problem arising from rounding errors. But it would not solve the problems of potentially wrong inclusion or exclusion.

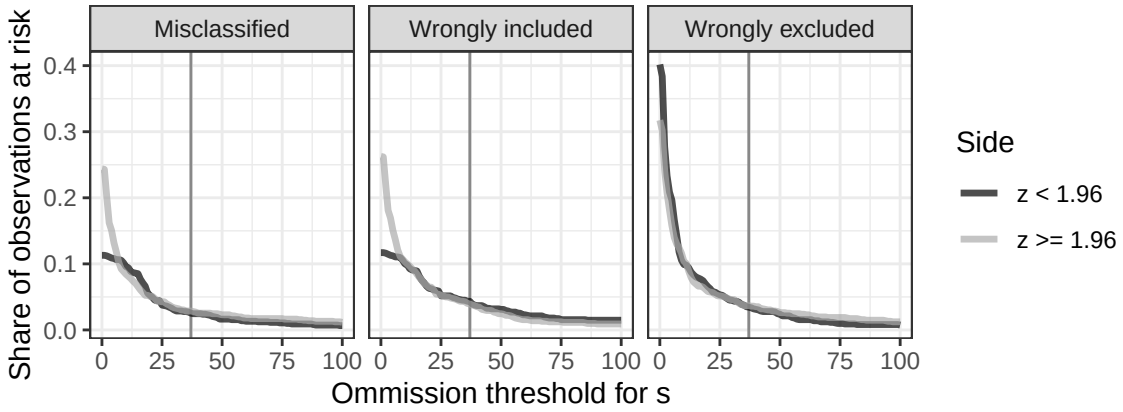

Figure 3: Share of observations at risk of misclassification, wrong inclusion or exclusion in a window with half-width $h = 0.1$.

Figure 3 illustrates the risks for a window with half-width $h = 0.1$. It shows, as function of the omission threshold for the significand s, the share of observations relative to the total observations in the window which are at risk of being misclassified or wrongly included, as well as the relative number of observations outside the window at risk of being wrongly excluded. Importantly, for very low omission thresholds, we have an asymmetry: More observations with reported $z \geq 1.96$ are at risk of being misclassified or wrongly included, which is driven to a large extent by the observations with $z = 2$. Likewise, more observations with $z < 1.96$ are at risk to be wrongly excluded. While there is still error potential for our threshold of $s = 37$ (gray vertical line), there are similarly many observations at risk on both sides of the $z = 1.96$ threshold.

In the right-hand side of Figure 1, we have applied our adjustment procedure to the pooled data. The discontinuity at $h = 0.04$ vanishes and no clear evidence for p-hacking or publication bias remains. The shares of significant z-statistics decrease for all window half-widths, they are mostly close to 50% and the confidence intervals always include the 50% level.

2.2 Comparison with derounding methods and Monte-Carlo study

While BCH do not adjust for rounding errors, Brodeur et. al. (2016) and Bruns et al. (2019) address the problem in similar studies with derounding procedures. Brodeur et. al (2016) assume that the unobserved digits of the reported coefficient μ and standard error σ are drawn from a uniform distribution and use such random draws to generate a single derounded data set. Bruns et al. (2019) reduce noise of this procedure by repeating this derounding procedure several times and report mean estimates. 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.[6]

To compare the different approaches, we perform a Monte-Carlo study with three scenarios. In the first scenario, the simulated true z-statistics $\tilde{z}_j$ are uniformly distributed on the interval $[0, 2 \cdot 1.96]$. In the second scenario, $\tilde{z}_j$ is simulated such that 35% of the observations are uniformly distributed on $[0, 1.96]$ and the 65% uniformly on $[1.96, 2 \cdot 1, 96]$, i.e. in each window around 1.96, we would expect 65% of the tests to be significant. The 3rd scenario just swaps these probabilities so that we would expect a 35% of the tests to be significant.

Except for the distribution of z-statistics, we follow in all scenarios closely the characteristics of BCH's pooled data set in the interval $[0, 2 \cdot 1.96]$, e.g. each simulated sample has the same number of observations (16,777). Each simulated observation j is linked to a randomly drawn (with replacement) observation i from BCH's data. The simulated true standard error $\tilde{\sigma}_j$ is set to a uniformly derounded version of the standard error σ_i of BCH's observation i. We then compute the simulated true coefficient $\tilde{\mu}_j = \tilde{z}_j \tilde{\sigma}_j$. We denote by μ_j and σ_j the simulated reported coefficients and standard errors, created by rounding $\tilde{\sigma}_j$ and $\tilde{\mu}_j$ to the same number of decimal places as observation i in BCH's data. Finally, $z_j = \mu_j / \sigma_j$ is the simulated reported z-statistic. For each simulated sample, we compare different approaches to deal with the rounding problem. We repeat the procedure to get 100,000 simulated samples and show the averaged results for the first two scenarios and a window with half-width $h = 0.2$ in Table 3. The

[6] 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.

results for the 3rd Scenario are in Table 8 in the Appendix A1.

Table 3: Results of Monte-Carlo simulations (Scenarios 1 and 2)

	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
Adjusting derounding for density of z-statistics										
True density	0.000	0.476	0.524	97.1%	0.011	0.002	0.630	0.675	96.5%	0.011
Estimate (bw=0.2)	0.000	0.477	0.524	96.9%	0.011	-0.015	0.612	0.658	77.7%	0.018
Estimate (bw=0.05)	0.000	0.477	0.524	94.6%	0.012	-0.006	0.621	0.666	91.6%	0.013

Not adjusting for rounding causes a substantial upward bias in all three scenarios and the simulated coverage probabilities of the 95% confidence intervals are only 10.2%, 38.8% and 0.6%, respectively. In contrast, all adjustment approaches work well in the first scenario where the null hypothesis of an equal share of significant and non-significant tests is satisfied.

Note that the 95% confidence intervals of the median uniform derounding approach even achieves an excessive coverage probability of 97.1%. For an intuition of this excess coverage, we have also run a simulation where we use the original z-statistics $\tilde{z}_j$ and "deround" them by just adding symmetric, uniformly distributed noise directly to each $\tilde{z}_j$ value, again computing the median results of 100 such "derounding" samples. The 95% confidence intervals from this procedure also have excess coverage in the first scenario. Since the noise is drawn from a uniform distribution just as the original $\tilde{z}_j$, it corrects for imbalances in $\tilde{z}_j$ that arise due to sampling variation. While the noise itself also has random imbalances, that randomness is damped by taking the median over several samples.

Our uniform derounding approach that takes medians over several derounding samples seems to have a similar effect. In contrast, there is no clear excess coverage when taking a single derounding sample as in Brodeur et al. (2016). Also in terms of the root mean squared error (RMSE) the median derounding procedure is slightly superior

to the other two adjustment procedures.

The flip side of both uniform derounding approaches is that they induce an attenuation bias if the null hypothesis of locally uniform distributed z-statistics is not satisfied. Consider the 2nd scenario in which we have a 65% probability that a $\tilde{z}_j$ is above 1.96. There the 95% confidence intervals of both derounding approaches have low coverage (66.8% and 69.0%) and are biased towards finding a 50% share of significant z-statistics. We find similar results for the 3rd scenario.

Our approach to omit observations whose reported standard error has a significand below 37 seems not to suffer from relevant systematic biases in the simulation study and its confidence intervals achieve approximately 95% coverage in all three scenarios. Thus if the goal is to get unbiased estimators, the omission approach seems preferable. On the other hand, the median uniform derounding approach has confidence intervals with higher coverage and a slightly lower RMSE if the null hypothesis is satisfied. That derounding approach may thus be preferable for conservative tests of the null hypothesis.

2.3 z-density adjusted derounding (ZDA)

The problem of the uniform derounding approaches in the 2nd and 3rd scenario is that they do not account for the fact that the original z-statistics are less likely to come from the region where the z statistics have lower density. We propose two derounding approaches that take this issue into account.

The three bottom rows of Table 3 refer to an approach that we call z-density adjusted (ZDA) derounding. It modifies the derounding procedure by adjusting the derounding draws for the (estimated) true density of the z-statistics. In the first row, we use the true density $f(z)$, which shall be normalized such that its maximum is 1. An adjusted derounding draw for observation i is chosen as follows. We first draw a uniformly derounded value $\hat{z}_i$ as before but randomly reject that draw with probability $1 - f(\hat{z}_i)$. If we reject a draw, we repeat the procedure until a drawn derounded value is accepted. This well known rejection sampling technique causes the adjusted derounded values $\hat{z}_i$ to be drawn from a distribution whose density is proportional to the product $f(z)$ and the density implied by the uniform derounding approach. Except for this adjustment, we proceed as in the described median uniform derounding approach. In the first scenario the adjustment has no effect since the z-statistics are uniformly distributed and thus no draw will be rejected. Yet, in the 2nd scenario the adjustment achieves again excess

coverage of 97.1% and an estimated bias close to zero.

These results shall only illustrate that the method would work under ideal conditions. Yet, outside Monte-Carlo simulations the true density $f(z)$ is obviously unknown. The final two rows consider a feasible variant of the procedure by replacing $f(z)$ with a kernel estimate $\hat{f}(z)$ of the normalized density. We estimate that density using the subsample of z-statistics whose standard error is reported with at least three significant digits. We only perform the z-density adjusted derounding for observations whose standard error has less than three significant digits and otherwise perform uniform derounding. That is because sampling errors when estimating $\hat{f}(z)$ may become more severe if the derounded values can only fall into a small window around $\hat{z}$.

We present the results for Gaussian kernels with two different bandwidths. For a larger bandwidth results are more similar to the uniform derounding approach: excessive coverage in the first scenario but under-performance and an attenuation bias in the 2nd and 3rd scenario. The smaller bandwidth reduces the attenuation bias and achieves in all three scenarios a slightly better RMSE than our omission approach. Yet, the confidence intervals have coverage below 95%.

2.4 Derounding by simulating rounding (DSR)

Our final approach to derounding is computationally most expensive. It constructs the simulation of derounded z-statistics for any target observation i by explicitly modeling and simulating the scaling and rounding process that could have led to the reported μ_i, σ_i and z_i. We base the rounding simulations on the subset of observations S whose standard error is reported with at least 3 significant digits. A simplified version of this approach corresponds of the following steps:

1. For each observation $j \in S$ we first rescale its standard error σ_j to the scale of the target observation. More precisely, we set it to $\tilde{\sigma}_{ji} = \sigma_i$.

2. We then rescale the estimated coefficient μ_j such that the z-statistic z_j remains unchanged, i.e. we set $\tilde{\mu}_{ji} = z_j \tilde{\sigma}_{ji}$.

3. We then round $\tilde{\sigma}_{ji}$ and $\tilde{\mu}_{ji}$ to the same number of decimal places as observation i and let z_{ij} be the reported z-statistic given this rounded value.

4. We collect the original z-statistics z_j from all observations where the rounded statistic z_{ij} is equal to the reported z-statistic z_i of our target observation.

5. We repeat steps 1-4 several times until we have collected at least 10,000 z-statistics z_j for target i. We also repeat the procedure for each target i whose z-statistic we want to deround. Afterwards we proceed as in our earlier derounding approaches but pick the derounding draw for each target observation i from the more than 10,000 collected z-statistics z_j.

To get a smoother distribution of the derounded z-statistic for each target observation z_i, our actual procedure modifies the simplified procedure above as follows. Instead of using the actual observations σ_i, μ_j, σ_j, we draw each time uniformly derounded values $\hat{\sigma}_i$, $\hat{\mu}_j$ and $\hat{\sigma}_j$. E.g. if $\sigma_i = 0.05$ we draw $\hat{\sigma}_i$ from a uniform distribution on the interval $[0.045; 0.055]$. Instead of z_j, we use $\hat{z}_j = \hat{\mu}_j/\hat{\sigma}_j$.

We will apply the derounding by simulating rounding (DSR) procedure on BCH's data set but do not perform a Monte-Carlo simulation of the approach, since it is too time consuming. Similar to the z-density adjusted derounding, we apply DSR derounding only for target observations whose standard error is reported with less than 3 significant digits and perform uniform derounding for the remaining observations.

3 Reanalyzing BCH's randomization and caliper tests

In this section we replicate BCH's randomization and caliper tests at the 5% significane threshold using the different approaches to adjust for (coarse) rounding. BCH compare the evidence for p-hacking between four different identification strategies: difference-in-difference (DID), instrumental variable (IV), randomized experiment (RCT) and regression discontinuity design (RDD). Table 4 shows that the share of observations with a z-statistic of exactly 2 differs substantially between the subsamples corresponding to the different identification strategies. In the smallest window studied by BCH, it ranges from only 16.6% for IV to 50.0% for DID. Correspondingly, adjusting for rounding errors affects in particular the DID results.

3.1 Randomization tests

Table 3 in BCH shows the share $\hat{\theta}$ of observations with $z \geq 1.96$ in the window $[1.96 - h; 1.96 + h]$ for different window sizes h and different identification strategies. It also presents the p-value of a one-sided binomial test with the null hypothesis $\theta \leq 0.5$.[7]

[7]To facilitate the comparison between so many coefficients, we decided to report significance stars in this working paper even though in other cases there are good arguments against it.

Table 4: Sample split by method

Method	n (total)	n ($z = 2$)	Share of observations with $z = 2$ in window with half-width $h = 0.05$
DID	5,853	115	50.0%
IV	5,170	28	16.6%
RCT	7,569	83	39.9%
RDD	3,148	34	43.0%

Table 5 compares the corresponding results for the subsample of DID tests data set for the different approaches of dealing with the rounding problem. Without adjustment for rounding, we find for all considered windows that the share $\hat{\theta}$ of tests of with $z \geq 1.96$ is significantly above 50% and ranges between 53.1% and 66.5% (with p-values between 0.000 and 0.031).[8] In contrast, when adjusting for rounding, $\hat{\theta}$ varies between 44.6% and 55.3% and is not significantly above 50% (at a significance level of 5%) for any combination of window size and adjustment approach (with p-values between 0.072 and 0.904). This means when adjusting for the rounding problem, the randomization tests no longer deliver statistical evidence for p-hacking or publication bias for DID studies.[9]

The results for the other subsamples are given in Appendix A3. For the IV subsample (Table 9) the adjustment for rounding has relatively little effect and we find for all windows sizes very similar effect sizes and p-values that are mostly significant. Tables 10 and 11 present the results for the RDD and RCT subsamples. Without adjustment one finds for small window half-widths shares $\hat{\theta}$ that are significantly above 50%. When adjusting for rounding the estimator $\hat{\theta}$ is not significantly above 50% for any window size.

3.2 Caliper tests

BCH proceed their analyses with so-called caliper tests. Again, all observations with z-statistics in a specified window around the $z = 1.96$ threshold are considered and

[8] Note that in BCH's original table these shares vary between 53.0% and 70.7%. The differences are due to the reasons explained in Footnote 1.

[9] One might argue that by destroying the significance stars in Table 5, the adjustment brought a new hope for the current state of empirics. See Lucas et al. (1977) for a related insight concerning star destruction.

Table 5: Randomization tests for DID subsample.

	(1) No adj.	(2) Omit	(3) Uniform	(4) ZDA	(5) DSR
Window half-width 0.05					
Proportion Significant	0.665***	0.446	0.52	0.513	0.531
(p-value)	(0.000)	(0.884)	(0.362)	(0.415)	(0.307)
No. obs.	230	101	132	134	99
Window half-width 0.075					
Proportion Significant	0.654***	0.486	0.532	0.531	0.547
(p-value)	(0.000)	(0.659)	(0.189)	(0.201)	(0.141)
No. obs.	292	148	211	213	156
Window half-width 0.1					
Proportion Significant	0.628***	0.507	0.541*	0.539	0.553*
(p-value)	(0.000)	(0.446)	(0.089)	(0.106)	(0.075)
No. obs.	382	217	282	287	206
Window half-width 0.2					
Proportion Significant	0.553***	0.469	0.508	0.505	0.511
(p-value)	(0.004)	(0.904)	(0.374)	(0.416)	(0.346)
No. obs.	636	397	549	563	411
Window half-width 0.3					
Proportion Significant	0.531**	0.478	0.505	0.503	0.507
(p-value)	(0.031)	(0.868)	(0.402)	(0.451)	(0.373)
No. obs.	930	584	794	812	602
Window half-width 0.4					
Proportion Significant	0.537***	0.497	0.509	0.511	0.516
(p-value)	(0.007)	(0.574)	(0.282)	(0.258)	(0.200)
No. obs.	1161	720	1029	1048	778
Window half-width 0.5					
Proportion Significant	0.534***	0.506	0.517	0.519*	0.516
(p-value)	(0.006)	(0.367)	(0.122)	(0.088)	(0.166)
No. obs.	1391	869	1245	1263	950

$^{*}\, p \leq 0.1,\ ^{**}\, p \leq 0.5,\ ^{***}\, p \leq 0.01$

probit regressions of the following form are performed:

$$Pr(Significant_i = 1) = \Phi(\alpha + X_i'\boldsymbol{\delta} + \gamma DID_i + \lambda IV_i + \phi RDD_i).$$

$Significant_i$ is a dummy variable indicating whether $z_i \geq 1.96$ and X_i is a vector of control variables, including author and article characteristics. Table 6 shows our replication results of the caliper tests when adjusting for rounding errors by omitting all observations with $s < 37$.

Table 6: Caliper tests using only observations with $s \geq 37$

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.056	0.045	0.029	0.028	0.024	-0.036
	(0.038)	(0.039)	(0.040)	(0.041)	(0.045)	(0.059)
IV	0.101***	0.100***	0.079**	0.084**	0.097**	0.098*
	(0.034)	(0.037)	(0.038)	(0.038)	(0.041)	(0.050)
RDD	0.094	0.082	0.069	0.069	0.074	0.038
	(0.063)	(0.061)	(0.058)	(0.058)	(0.060)	(0.073)
Top 5		-0.028	-0.021			
		(0.054)	(0.110)			
Year = 2018		0.006	0.013	0.022	-0.004	0.022
		(0.033)	(0.033)	(0.034)	(0.037)	(0.041)
Experience		-0.008	-0.013	-0.013	-0.011	0.000
		(0.008)	(0.008)	(0.008)	(0.010)	(0.011)
Experience^2		0.002	0.017	0.020	0.020	-0.005
		(0.021)	(0.022)	(0.022)	(0.026)	(0.032)
Top institution		-0.001	0.003	0.002	-0.040	-0.078
		(0.054)	(0.053)	(0.052)	(0.059)	(0.069)
Top PhD institution		0.014	-0.013	-0.013	0.058	0.143**
		(0.045)	(0.045)	(0.046)	(0.050)	(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	3,558	3,558	3,558	3,558	2,626	1,585
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	.47	.47	.47	.47	.48	.49

Marginal effects; Standard errors in parentheses

(d) for discrete change of dummy variable from 0 to 1

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: 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. * $p \leq 0.1$, ** $p \leq 0.5$, *** $p \leq 0.01$.

E.g. the regression estimates in the first column indicate that in the window $[1.96 - 0.5; 1.96 + 0.5]$ tests from IV studies are 10.1 percentage points more likely to be significant than tests from RCT studies of which 46.9% are significant in that window. Overall, the estimates for IV are very similar to those of BCH and remain

significant when performing our adjustment for rounding errors. In contrast, effect sizes for DID substantially reduce with our adjustment and, contrasting the findings of BCH, are not significant in any specification. So after adjustment for rounding errors, the caliper tests provide no more evidence for p-hacking of DID studies at the 5% threshold than the randomization tests.

4 Further Reanalysis and Issues Unrelated to Rounding

In this section we reanalyze further results of BCH and point out issues mainly unrelated to rounding.

4.1 Distribution of z-statistics

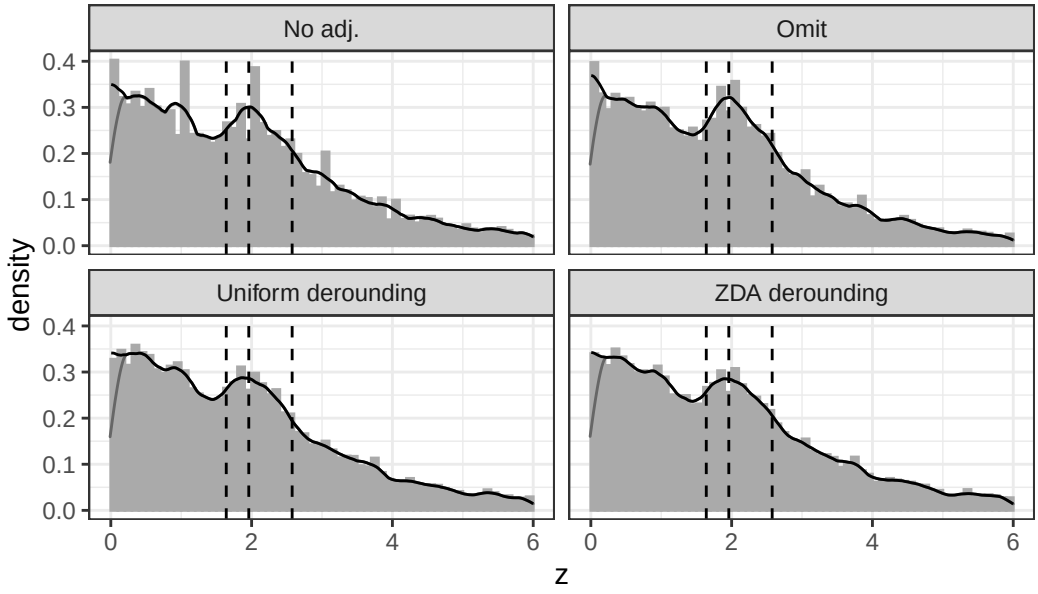

Figure 4: Pooled data. Distribution of z-statistics for different adjustment methods.

Notes: For both derounding approaches a single random draw of z-statistics is taken. The dashed vertical lines indicate the critical z-thresholds 1.645, 1.96 and 2.576 at the 10%, 5% and 1% significance levels, respectively. The histograms have bin size 0.05. The black line is an density estimate with a Epanechnikov kernel with bandwidth 0.1. We adjust density estimates at the left margin by assuming that the distribution is symmetric for positive and negative z-statistics. The dark gray line on the left shows the standard density estimate without that adjustment, which is biased towards zero at the left margin.

Figure 4 shows the distribution of z-statistics of the pooled data set for different approaches to the rounding problem. Compared to no adjustment, the main effect of the adjustment procedures is to remove the bunching at particular z-statistics; most importantly the bunching at $z = 2$ just right of the 5% significance threshold that caused the substantial imbalance in the randomization tests and the caliper tests. The general shape of the kernel density estimate with two humps looks relatively similar across all approaches, including the no adjustment case. Intuitively, kernel density estimates themselves smooth out the bunching points and thus have some analogous effects to derounding. Appendix A2 shows similar results also in the subsamples separated by identification strategy.

Note that the default kernel density estimator (dark gray line on left) used by BCH is biased towards zero at the left margin. We use a corrected density estimator (black line) that is based on the assumption that original z-statistics where distributed symmetrically around 0. We see that the corrected kernel density is much more in line with the histograms close to the $z = 0$ margin. While BCH labelled the form of original kernel density estimator as camel shaped this label fits less obviously with the corrected density estimator where the first hump peaks at zero.

4.2 Excess Test Statistics

In their Section 3C, BCH hypothesize that absent publication bias and p-hacking, the distribution of z-statistics would follow for each causal identification strategy a non-central t-distribution truncated at $z = 0$. They calibrate, separately for each subsample, the degrees of freedom and non-centrality parameters of those t-distributions by matching only the tails of the empirical distribution with $z > 5$.

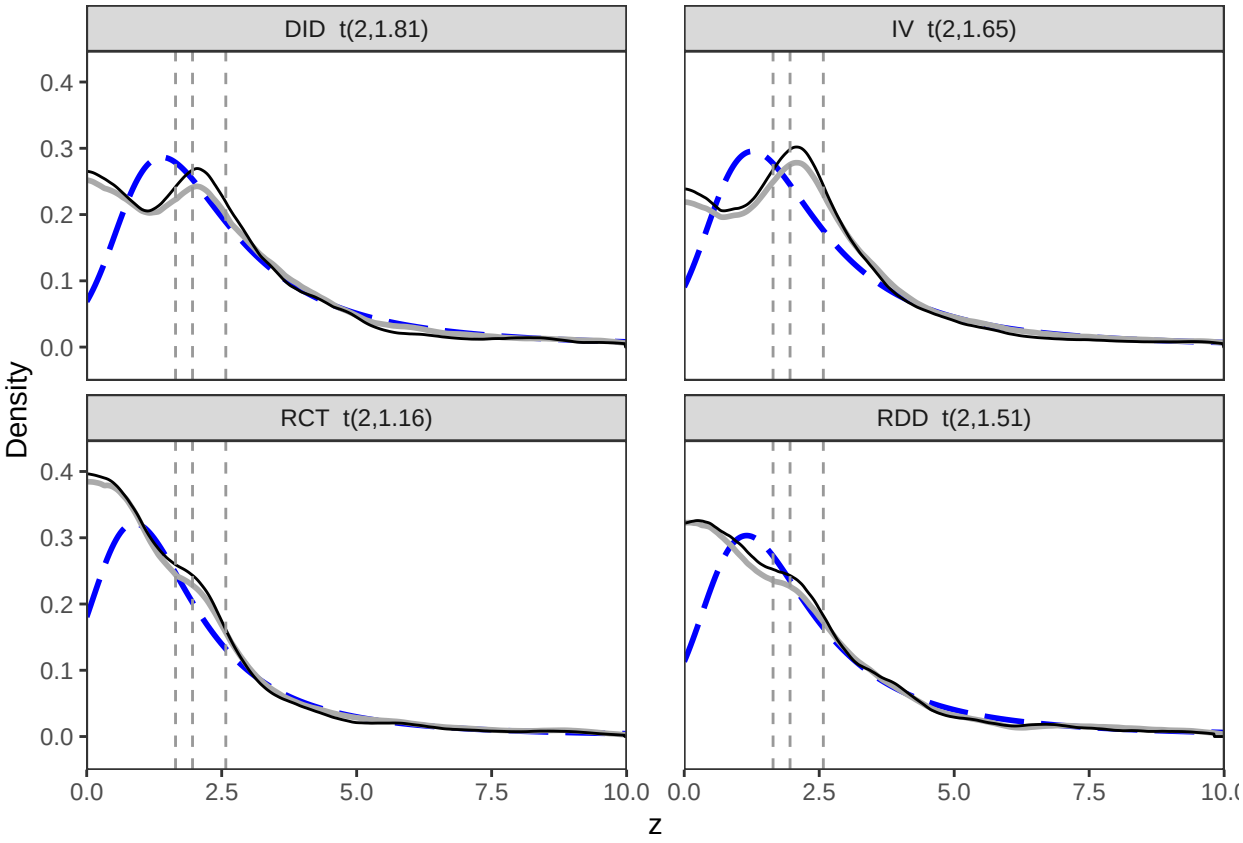

Figure 5: Excess test statistic plots corresponding to Figure 4 in BCH.

Notes: The dashed blue line shows the density of BCH's originally estimated non-central t-distribution for each subsample. The black line is the empirical kernel density estimate using omission as rounding adjustment and the gray line for the case of no rounding adjustment. As in Figure 4 we use adjusted kernel estimates that prevent the downward bias at the left-side margin.

Figure 5 corresponds to Figure 4 in BCH and compares these t-distributions with the kernel estimates of the empirical densities.[10] Whether we adjust for rounding (black line) or take the unadjusted z-statistics (gray line) leads visually only to small differences. BCH suggest that the differences between the empirical densities and the t-distribution indicate distortions arising from publication bias or p-hacking. They derive quantitative distortion measures by comparing the probability masses in certain intervals such as $[1.96, 2.58]$. Given that the unsmoothed empirical distributions are used for those measures, adjustment for rounding errors may have a somewhat stronger

[10]When comparing these t-distributions with the empirical densities, BCH showed in their Figure 4 by mistake the densities of the t-distributions without accounting for the truncation at zero. Figure 5 shows the corrected densities.

effect than Figure 5 suggests. We do not argue though that rounding errors importantly change the qualitative insights from these exercises. Yet, we see other problems.

The calibration problem

BCH motivate their calibration as follows (p. 3650): *"We assume that the observed test statistic distribution above $z = 5$ should be free of p-hacking or publication bias—the incentives to p-hack in a range so far above the traditional significance thresholds are plausibly zero."*

However, even if publication bias only reduces the publication probability for small absolute z-statistics, it will also affect the observed density for z above 5 because the total area under the density function must always be one.

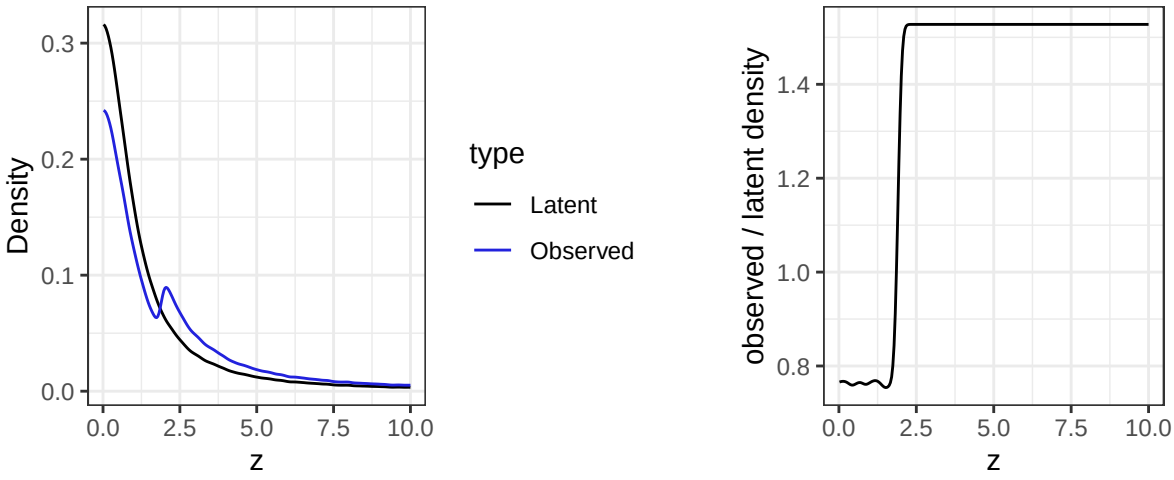

Figure 6: Latent vs. observed densities in a model with publication bias for $z \leq 1.9$.

Notes: We consider a model in which the latent distribution of z-statistics absent publication bias follows a t-distribution with one degree of freedom. The observed distribution of z-statistics is subject to a publication bias such that observations with $abs(z) \leq 1.9$ only have a 50% probability to be published. The left panel shows the densities of the latent and observed distribution and the right panel the density ratios.

Figure 6 illustrates this point with a simple simulated data set. The latent distribution of z-statistics absent publication bias follows a t-distribution with one degree of freedom. The observed distribution of z-statistics is subject to a publication bias such that observations with $abs(z) \leq 1.9$ only have a 50% probability to be published. For absolute z-statistics below 1.9, this publication bias leads to a lower density in the

observed z-statistics compared to the latent distribution. Yet, on the other hand, for absolute z-statistics above 1.9, it also yields a more than 50% higher density for the observed z-statistics compared to the latent distribution. Hence, fitting a non-central t-distribution by matching the tails of the observed distribution for $z \geq 5$ would not recover the original distribution of z-statistics. More generally, we can establish

Proposition 1. *Let $F^*(z)$ be a distribution function of the 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 $z_2 \to \infty$. □

Assumed benchmark distributions have smaller probability mass for small z-statistics than observed distributions

Compared to the empirical distribution of z-statistics, the latent distributions proposed by BCH have much smaller densities for z-statistics close to 0. There seems no plausible explanation why p-hacking or publication bias should increase the density of z-statistics close to zero, we would rather expect the opposite. This observation was not evident in BCH since they used as biased kernel density estimator that wrongfully suggested also low empirical density of z-statistics close to zero.

4.3 Approach of Andrews and Kasy (2019)

Andrews and Kasy (2019) explore two approaches to estimate the relative publication probabilities conditional on a z-statistic and the benchmark distribution of z-statistics

absent publication bias. The first approach relies on an appropriate subset of studies for which p-hacking or publication bias is unlikely, e.g. studies from systematic replication projects. The second approach crucially relies on the identifying assumption that in the unobserved latent distribution absent publication bias, the standard error σ_i and estimated coefficient μ_i are independently distributed from each other.

BCH adopt this second approach in section 3D assuming the benchmark distribution for absolute z-statistics is a generalized t-distribution. First, note that many of our earlier remarks concerning the assumption of such a uni-modal distribution for absolute z-statistics still apply. Yet, perhaps more importantly, we do not believe that it is valid to use this approach by Andrews and Kasy (2019) for the given data set, because the independence assumption between σ_i and μ_i seems strongly violated.

For a statistical test, we compute for all observations the weighted correlation between $\log \sigma_i$ and $\log abs(\mu_i)$ using as weights the inverse of the estimated publication probabilities. Under the null hypothesis that all assumptions of the chosen implementation of the Andrews and Kasy (2019) approach are satisfied, this inverse probability weighting allows to recover the correlation in the unobserved latent distribution of tests if no publication bias was present.[11] Table 7 shows the computed correlations and bootstrapped confidence intervals for different subsamples.

[11]We thank Isaiah Andrews who proposed the idea for this inverse probability weighting approach and pointed out a problem of an earlier idea of ours to test the independence assumption.

Table 7: Specification test for Andrews and Kasy (2019) appoach

	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]

Notes: The table shows the inverse probability weighted correlations between $\log \mu$ and $\log \sigma$ as explained in Section 4.3. 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.

The correlations range from 0.89 to 0.92 and the 95% confidence intervals from 0.88 to 0.93. These results suggest that the crucial independence assumption of Andrews and Kasy (2019) is strongly violated in BCH's data set.

Our intuition for that correlation is that a lot of variation in σ_i and μ_i is driven by the scaling of the dependent and explanatory variables in the regressions. For example, if we rescaled all explanatory variables in a regression by multiplying them with a factor $1/m$, then the corresponding μ_i and σ_i would both change by the factor m. This can lead to a strong positive correlation of $\log \sigma_i$ and $\log abs(\mu_i)$ even absent publication bias.

5 Concluding discussion

5.1 Interpreting the 2nd hump

Even after adjusting for rounding, the empirical distribution of z-statistics robustly exhibits two humps, one at zero and one loosely around $z = 2$. Is this 2nd hump already evidence for p-hacking or publication bias? Not necessarily. An alternative explanation could be that the latent distribution is a mixture distribution arising from differently targeted research questions. For example, some studies are likely to refine previous re-

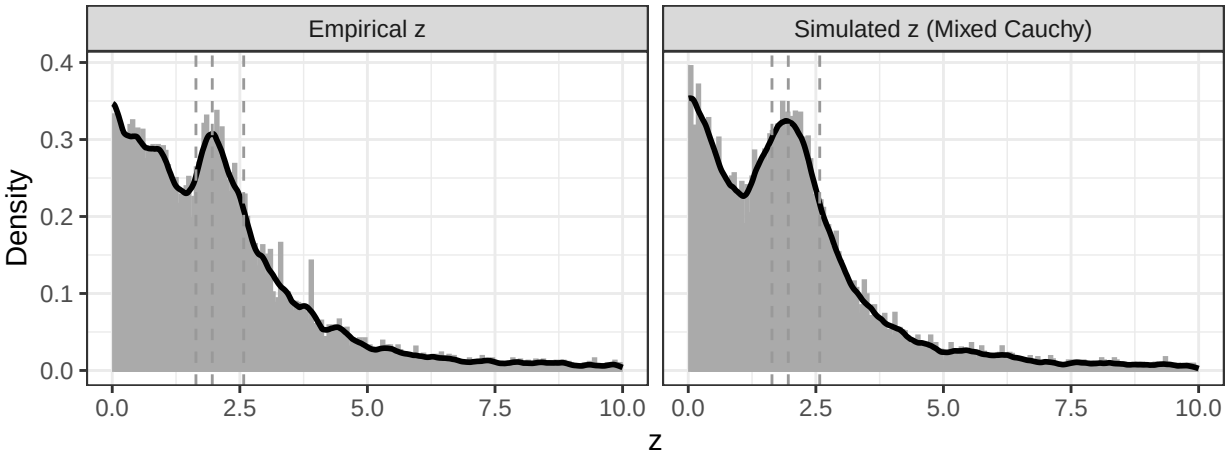

Figure 7: Comparing empirical z-statistics (pooled data with $s \geq 37$) with random draws of mixed Cauchy distributions.

search that found significant effects, while other research questions are more exploratory without a strong prior that actual effects should be present. For a numerical illustration, we pick 5,000 random draws from an equal mixture of three Cauchy distributions all with scale parameter 0.8: one distribution has a location parameter of 0 and shall correspond to exploratory research and the other two have location parameters -2 and 2 and shall correspond to more targeted research.

Figure 7 shows that the resulting distribution of absolute z-statistics is very similar to the empirical distribution in the pooled data (excluding observations with $s < 37$). This simple model of differently focused research can similarly generate two humps, even though p-values are not hacked and the publication probability does not depend on the resulting z-statistic of any research project. Of course, the 2nd hump could also be driven to some extent or even completely by publication bias or p-hacking. In our view, it is very hard to find a convincing route to disentangle the sources of the hump using BCH's data set.

One may argue that allowing mixture distributions as latent distributions just grants too much flexibility. Note, however, that in their excess statistics approach BCH assume that the parameters of the latent t-distributions vary across empirical methods. This implicitly also assumes that the z-statistics for the pooled data are drawn from a mixture of different non-central t-distributions.

Moreover, note that any latent distribution with a peak strictly above zero for the absolute z-statistics implies, similar to mixture distributions, more than one peak for

the distribution of all negative and positive z-statistics if we assume z-statistics are symmetrically distributed around 0.

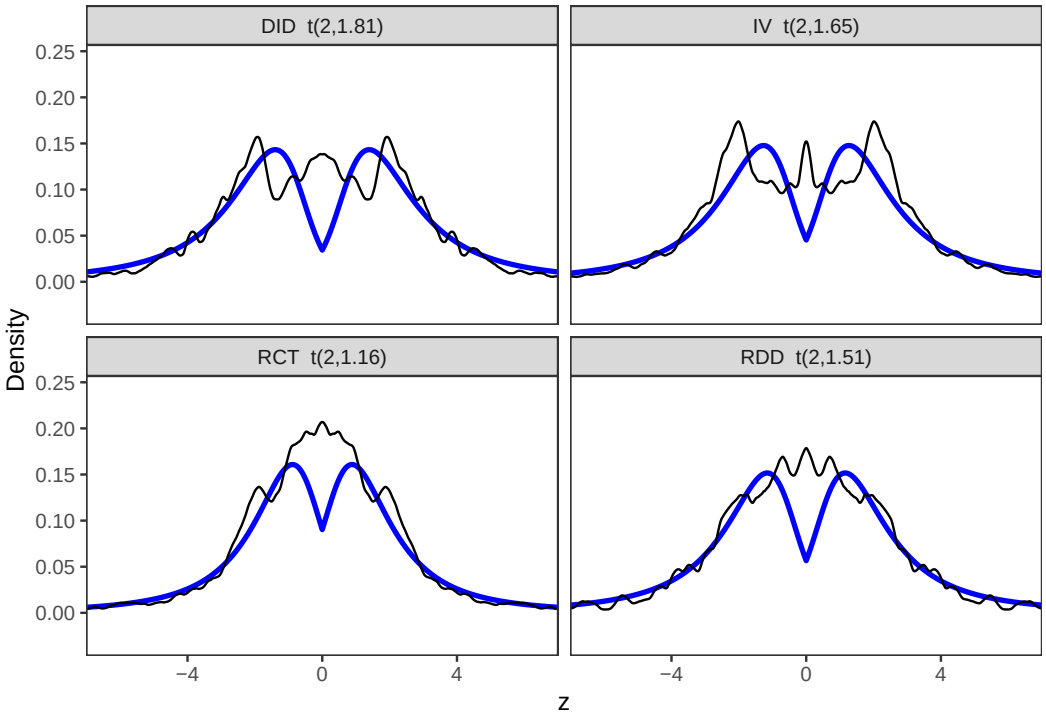

Figure 8: Kernel density estimators for empirical distributions (black line) vs densities of BCH's latent distributions (blue line) assuming that z-statistics are symmetrically distributed around 0.

For an illustration consider Figure 8. It compares BCH's calibrated latent distributions with the empirical densities of z-statistics under the assumption that z-statistics are symmetrically distributed around 0. Functional form assumptions are to some degree subjective, but in our view a mixture distribution that has an additional hump at $z = 0$ seems a more plausible latent distribution than the considered distributions with a trough at $z = 0$.

5.2 Randomization tests (also at the 10% and 1% threshold) and the 2nd hump

Tables 12 and 13 in the appendix show the results of the randomization tests at the 1% and 10% significance thresholds. At the 10% threshold we find significantly more than 50% of tests above the threshold for all window sizes for IV and the pooled

data, and for larger window sizes also for DID and selectively for RCT. As Figure 4 shows, densities of z-statistics generally slope upwards around this threshold. In contrast, at the 1% threshold, the densities slope downwards and the share of tests with $z \geq 2.576$ is significantly below 50% for many window half-widths. So the interpretation of the randomization tests at these thresholds is not straightforward as it depends on interpretation of the hump whose slope affects the results. The randomization tests with smaller window sizes at the 5% threshold seem less affected by the interpretation of the hump since this threshold lies in all subsamples close to the peak of the 2nd hump where the assumption of a locally flat slope absent p-hacking and publication bias seems more plausible.

5.3 Comparing densities with RCT densities

BCH also discuss the idea, albeit noting that they do not prefer it, to use the density for the RCT subsample as a benchmark for the densities of the DID, IV and RDD subsamples. Figure 9 shows this comparison. We see how RCT publications have relatively more insignificant z-statistics than all other three methods, and the differences are particularly pronounced for IV and DID.

Since the total density must be balanced, we find a relatively higher share of significant tests for DID and IV studies than for RCT. As the log ratios of the densities in the lower panel of Figure 9 show, the relatively higher density of significant results is not especially focused on the regions around the significance thresholds, where one would mainly expect p-hacking but peaks beyond the 1% significance threshold. BCH's interpretation is that this latter observation shows a weakness of this approach. Nevertheless, we find these findings very interesting, in particular, when considering interpretations beyond p-hacking.

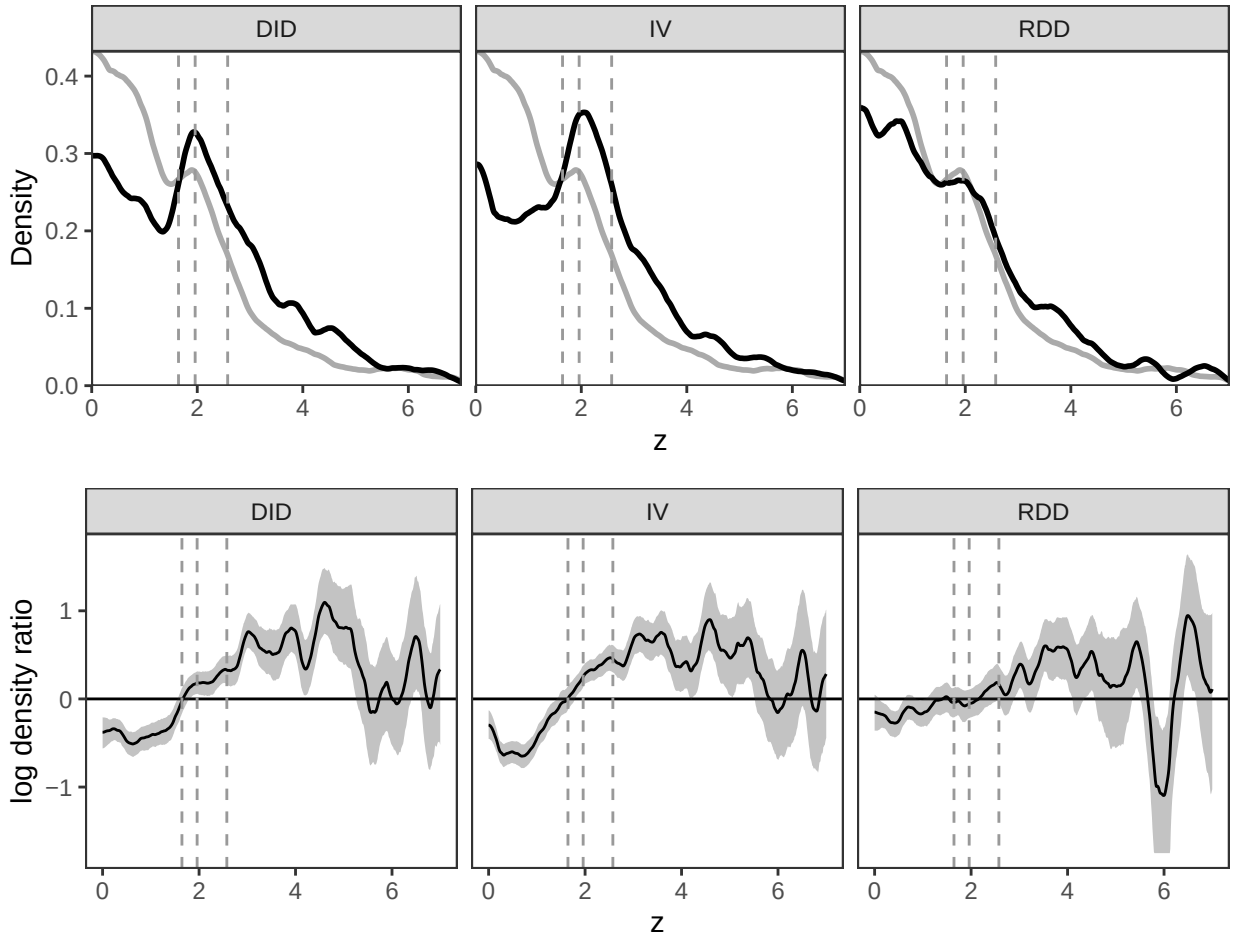

Figure 9: Comparison with RCT density

Notes: The upper panel shows the kernel estimates of densities of the z-statistic for RCT (gray) and the corresponding other subsample (black) considering all observations with $s \geq 37$. The lower panel shows the logarithm of the ratio of both estimated densities with the RCT density in the denominator with bootstrapped 95% CI.

One interpretation of Figure 9 is that it mainly shows a relative publication bias with lower publication hurdles for insignificant results in RCT than in quasi-experimental studies. Whether or not such heterogeneous publication hurdles might be warranted, e.g. if the potential for p-hacking differs between RCT and non-RCT, is an interesting discussion but beyond the scope of this paper.

REFERENCES

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and correction for publication bias." *American Economic Review* 109 (8): 2766-94.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review*, 110 (11): 3634-60.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics*, 8 (1): 1-32.

Bruns, Stephan B, Igor Asanov, Rasmus Bode, Melanie Dunger, Christoph Funk, Sherif M. Hassan, Julia Hauschildt, et al. 2019. "Errors and Biases in Reported Significance Levels: Evidence from Innovation Research." *Research Policy* 48 (9): 103796.

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", mimeo.

Lucas, George et al. 1977. "Star Wars: Episode IV - A New Hope"

Pütz, Peter, and Stephan B. Bruns. 2021 "The (Non-) Significance Of Reporting Errors In Economics: Evidence From Three Top Journals." *Journal of Economic Surveys* 35.1 (2021): 348-373.

Appendix: Additional Figures and Tables

A1: Results of third scenario of Monte-Carlo Simulation

Table 8: Results of Monte-Carlo simulations (Scenario 3)

	Share significant: 35%				
Approach	Est. Bias	95% CI		Cover-age	RMSE
No adjustment	0.052	0.379	0.425	0.6%	0.053
Omit $s < 37$	0.001	0.324	0.379	94.9%	0.014
Derounding assuming uniform distribution of unobserved digits					
Single sample	0.017	0.345	0.390	68.3%	0.021
Median	0.017	0.345	0.390	70.5%	0.020
Adjusting derounding for density of z-statistics					
True density	-0.002	0.325	0.370	96.7%	0.011
Estimate (bw=0.2)	0.015	0.342	0.388	76.7%	0.019
Estimate (bw=0.05)	0.007	0.334	0.379	91.0%	0.013

A2: z-density distributions

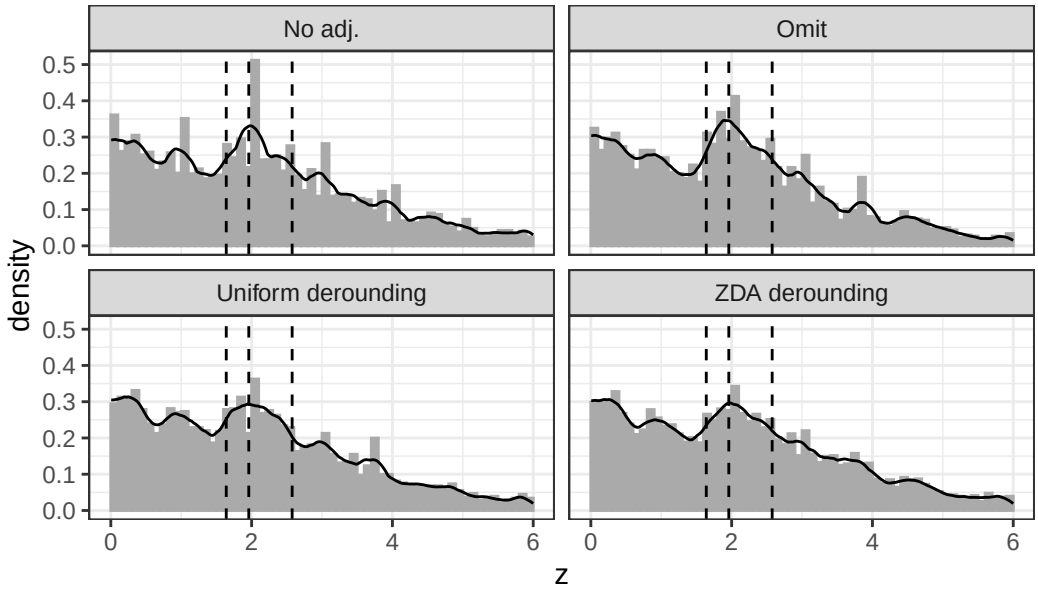

Figure 10: DID. Distribution of z-statistics for different adjustment methods. (For both derounding approaches a single random draw of z-statistics is taken).

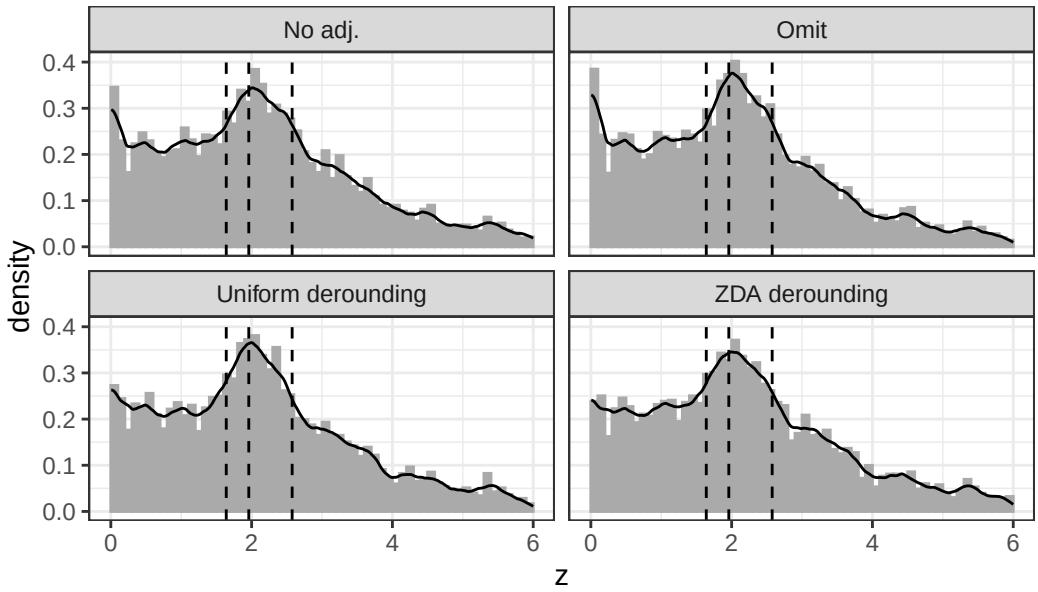

Figure 11: IV. Distribution of z-statistics for different adjustment methods. (For both derounding approaches a single random draw of z-statistics is taken).

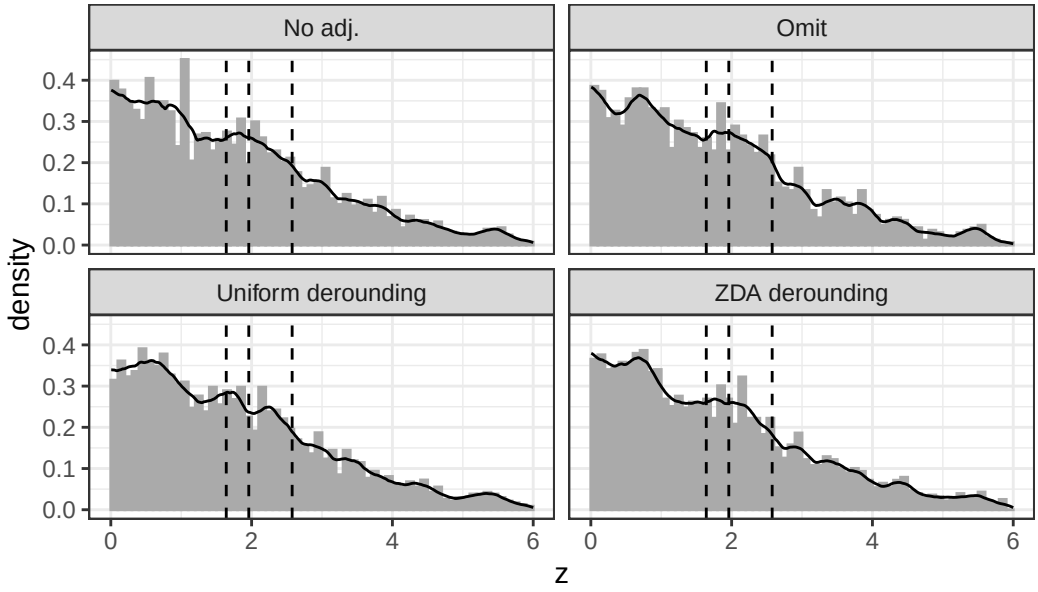

Figure 12: RDD. Distribution of z-statistics for different adjustment methods. (For both derounding approaches a single random draw of z-statistics is taken).

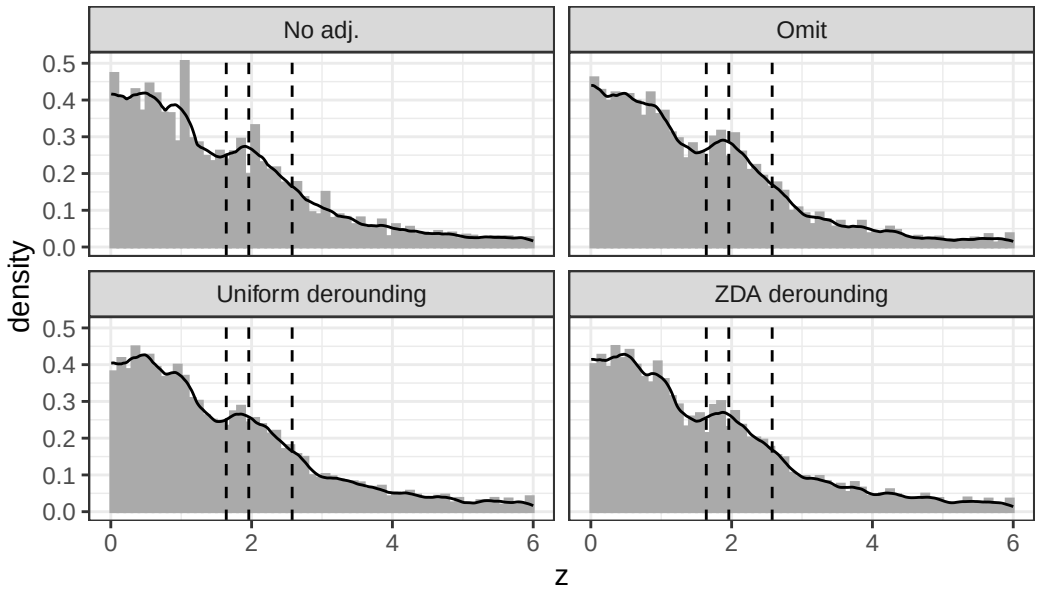

Figure 13: RCT. Distribution of z-statistics for different adjustment methods. (For both derounding approaches a single random draw of z-statistics is taken).

A3: Further Randomization Tests

Table 9: Randomization tests for IV, 5% significance threshold ($z = 1.96$)

	(1) No adj.	(2) Omit	(3) Uniform	(4) ZDA	(5) DSR
Window half-width 0.05					
Proportion Significant	0.609***	0.581**	0.547	0.55	0.579*
(p-value)	(0.003)	(0.036)	(0.137)	(0.122)	(0.051)
No. obs.	169	136	153	156	118
Window half-width 0.075					
Proportion Significant	0.596***	0.591***	0.562**	0.562**	0.593***
(p-value)	(0.002)	(0.006)	(0.036)	(0.033)	(0.008)
No. obs.	240	198	231	234	175
Window half-width 0.1					
Proportion Significant	0.57***	0.56**	0.549**	0.551**	0.571**
(p-value)	(0.008)	(0.029)	(0.047)	(0.042)	(0.018)
No. obs.	316	266	307	311	231
Window half-width 0.2					
Proportion Significant	0.539**	0.545**	0.533*	0.538**	0.546**
(p-value)	(0.029)	(0.023)	(0.058)	(0.036)	(0.027)
No. obs.	616	506	587	594	456
Window half-width 0.3					
Proportion Significant	0.524*	0.53*	0.521	0.524*	0.526*
(p-value)	(0.073)	(0.056)	(0.111)	(0.078)	(0.095)
No. obs.	919	736	867	875	683
Window half-width 0.4					
Proportion Significant	0.533**	0.539***	0.524*	0.528**	0.52
(p-value)	(0.013)	(0.009)	(0.055)	(0.033)	(0.127)
No. obs.	1165	929	1106	1116	861
Window half-width 0.5					
Proportion Significant	0.54***	0.548***	0.533***	0.537***	0.53**
(p-value)	(0.001)	(0.001)	(0.008)	(0.004)	(0.031)
No. obs.	1408	1123	1336	1344	1030

Notes: One sided p-values. * $p \leq 0.1$, ** $p \leq 0.5$, *** $p \leq 0.01$

Table 10: Randomization tests for RDD, 5% significance threshold ($z = 1.96$)

	(1) No adj.	(2) Omit	(3) Uniform	(4) ZDA	(5) DSR
Window half-width 0.05					
Proportion Significant	0.62**	0.395	0.442	0.444	0.4
(p-value)	(0.021)	(0.928)	(0.859)	(0.847)	(0.939)
No. obs.	79	38	68	66	49
Window half-width 0.075					
Proportion Significant	0.553	0.453	0.472	0.462	0.461
(p-value)	(0.151)	(0.809)	(0.752)	(0.812)	(0.787)
No. obs.	114	64	105	106	77
Window half-width 0.1					
Proportion Significant	0.535	0.476	0.468	0.46	0.468
(p-value)	(0.225)	(0.707)	(0.800)	(0.856)	(0.778)
No. obs.	142	84	142	143	103
Window half-width 0.2					
Proportion Significant	0.486	0.459	0.466	0.463	0.47
(p-value)	(0.700)	(0.882)	(0.896)	(0.908)	(0.830)
No. obs.	294	183	298	301	220
Window half-width 0.3					
Proportion Significant	0.48	0.491	0.48	0.484	0.482
(p-value)	(0.814)	(0.644)	(0.816)	(0.766)	(0.762)
No. obs.	452	265	449	451	332
Window half-width 0.4					
Proportion Significant	0.478	0.469	0.469	0.473	0.469
(p-value)	(0.860)	(0.890)	(0.939)	(0.914)	(0.909)
No. obs.	579	352	583	587	440
Window half-width 0.5					
Proportion Significant	0.475	0.48	0.461	0.465	0.456
(p-value)	(0.918)	(0.807)	(0.982)	(0.972)	(0.981)
No. obs.	708	431	705	710	529

Notes: One sided p-values. * $p \leq 0.1$, ** $p \leq 0.5$, *** $p \leq 0.01$

Table 11: Randomization tests for RCT subsample, 5% significance threshold ($z = 1.96$)

	(1) No adj.	(2) Omit	(3) Uniform	(4) ZDA	(5) DSR
Window half-width 0.05					
Proportion Significant	0.639***	0.5	0.497	0.5	0.528
(p-value)	(0.000)	(0.538)	(0.565)	(0.532)	(0.297)
No. obs.	208	110	154	156	120
Window half-width 0.075					
Proportion Significant	0.567**	0.494	0.487	0.488	0.487
(p-value)	(0.012)	(0.588)	(0.683)	(0.671)	(0.668)
No. obs.	298	180	246	252	194
Window half-width 0.1					
Proportion Significant	0.558**	0.51	0.504	0.5	0.523
(p-value)	(0.014)	(0.400)	(0.457)	(0.521)	(0.246)
No. obs.	382	247	339	347	259
Window half-width 0.2					
Proportion Significant	0.501	0.485	0.478	0.478	0.477
(p-value)	(0.485)	(0.763)	(0.887)	(0.889)	(0.874)
No. obs.	746	501	692	708	554
Window half-width 0.3					
Proportion Significant	0.497	0.477	0.471	0.469	0.469
(p-value)	(0.596)	(0.903)	(0.971)	(0.980)	(0.965)
No. obs.	1087	724	1010	1023	807
Window half-width 0.4					
Proportion Significant	0.493	0.477	0.471	0.471	0.461
(p-value)	(0.712)	(0.926)	(0.983)	(0.984)	(0.994)
No. obs.	1410	927	1308	1318	1028
Window half-width 0.5					
Proportion Significant	0.479	0.469	0.464	0.466	0.461
(p-value)	(0.959)	(0.984)	(0.998)	(0.997)	(0.997)
No. obs.	1715	1135	1605	1619	1260

Notes: One sided p-values. * $p \leq 0.1$, ** $p \leq 0.5$, *** $p \leq 0.01$

Table 12: Randomization tests at 10% significance threshold ($z = 1.645$)

	(1) ALL	(2) DID	(3) IV	(4) RCT	(5) RDD
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)
No. obs.	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)
No. obs.	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)
No. obs.	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)
No. obs.	1372	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)
No. obs.	2036	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)
No. obs.	2767	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)
No. obs.	3462	781	1023	446	1212

Notes: Observations with $s < 37$ are omitted to adjust for rounding errors. One sided p-values.
* $p \leq 0.1$, ** $p \leq 0.5$, *** $p \leq 0.01$

Table 13: Randomization tests at 1% significance threshold ($z = 2.576$)

	(1) ALL	(2) DID	(3) IV	(4) RCT	(5) RDD
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)
No. obs.	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)
No. obs.	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)
No. obs.	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)
No. obs.	1087	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)
No. obs.	1598	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)
No. obs.	2130	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)
No. obs.	2700	695	924	333	748

Notes: Observations with $s < 37$ are omitted to adjust for rounding errors. Here we show two sided p-values to indicate that those tests show significantly smaller shares of tests below the z=2.576 threshold. * $p \leq 0.1$, ** $p \leq 0.5$, *** $p \leq 0.01$