

Stommes, Drew; Aronow, P. M.; Sävje, Fredrik

**Working Paper**

## On the Reliability of Published Findings Using the Regression Discontinuity Design in Political Science

I4R Discussion Paper Series, No. 22

**Provided in Cooperation with:**  
The Institute for Replication (I4R)

*Suggested Citation:* Stommes, Drew; Aronow, P. M.; Sävje, Fredrik (2023) : On the Reliability of Published Findings Using the Regression Discontinuity Design in Political Science, I4R Discussion Paper Series, No. 22, Institute for Replication (I4R), s.l.

This Version is available at:  
<https://hdl.handle.net/10419/269243>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



No. 22  
I4R DISCUSSION PAPER SERIES

# **On the Reliability of Published Findings Using the Regression Discontinuity Design in Political Science**

Drew Stommes

P.M. Aronow

Fredrik Sävje

**March 2023**

## I4R DISCUSSION PAPER SERIES

I4R DP No. 22

# On the Reliability of Published Findings Using the Regression Discontinuity Design in Political Science

**Drew Stommes<sup>1</sup>, P.M. Aronow<sup>2</sup>, Fredrik Sävje<sup>3</sup>**

*<sup>1</sup>Yale University, Dept. of Political Science, New Haven/USA*

*<sup>2</sup>Yale University, Dept. of Political Science, Dept. of Statistics and Data Science, Dept. of Economics, and Yale School of Public Health, New Haven/USA*

*<sup>3</sup>Yale University, Dept. of Political Science and Yale School of Public Health, Dept. of Biostatistics, New Haven/USA*

MARCH 2023

Any opinions in this paper are those of the author(s) and not those of the Institute for Replication (I4R). Research published in this series may include views on policy, but I4R takes no institutional policy positions.

I4R Discussion Papers are research papers of the Institute for Replication which are widely circulated to promote replications and meta-scientific work in the social sciences. Provided in cooperation with EconStor, a service of the [ZBW – Leibniz Information Centre for Economics](#), and [RWI – Leibniz Institute for Economic Research](#), I4R Discussion Papers are among others listed in RePEc (see IDEAS, EconPapers). Complete list of all I4R DPs - downloadable for free at the I4R website.

I4R Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

### Editors

**Abel Brodeur**  
*University of Ottawa*

**Anna Dreber**  
*Stockholm School of Economics*

**Jörg Ankel-Peters**  
*RWI – Leibniz Institute for Economic Research*

# On the reliability of published findings using the regression discontinuity design in political science

Drew Stommes<sup>1</sup>, P. M. Aronow<sup>1,2,3,4</sup>, and Fredrik Sävje<sup>1,2</sup>

<sup>1</sup>*Department of Political Science, Yale University*

<sup>2</sup>*Department of Statistics & Data Science, Yale University*

<sup>3</sup>*Department of Economics, Yale University*

<sup>4</sup>*Department of Biostatistics, Yale School of Public Health*

March 12, 2023

## Abstract

The regression discontinuity (RD) design offers identification of causal effects under weak assumptions, earning it a position as a standard method in modern political science research. But identification does not necessarily imply that causal effects can be estimated accurately with limited data. In this paper, we highlight that estimation under the RD design involves serious statistical challenges and investigate how these challenges manifest themselves in the empirical literature in political science. We collect all RD-based findings published in top political science journals in the period 2009–2018. The distribution of published results exhibits pathological features; estimates tend to bunch just above the conventional level of statistical significance. A reanalysis of all studies with available data suggests that researcher discretion is not a major driver of these features. However, researchers tend to use inappropriate methods for inference, rendering standard errors artificially small. A retrospective power analysis reveals that most of these studies were underpowered to detect all but large effects. The issues we uncover, combined with well-documented selection pressures in academic publishing, cause concern that many published findings using the RD design may be exaggerated.

---

We thank Matias Cattaneo, Alex Coppock, Andrew Harris, Greg Huber, Will Hunt, Austin Jang, Josh Kalla, Winston Lin, John Marshall, Lilla Orr, Tara Slough, Jessica Sun, two anonymous reviewers, and the editor, Abel Brodeur, for helpful comments and discussions.

# 1 Introduction

The regression discontinuity (RD) design requires relatively weak assumptions for identification of causal effects (Cattaneo, Titiunik, & Vazquez-Bare, 2020). As a result, the method has been widely used in modern empirical political science, and its popularity continues to grow. Prominent studies examine the effect of incumbency and control over media (Boas & Hidalgo, 2011), and the effect of electing extremist candidates in primaries (Hall, 2015). Because the requisite identifying assumptions are weak, researchers may surmise that all empirical challenges faced when using the method are easily solved. In this paper, we demonstrate that this view is largely unfounded, highlighting the serious statistical challenges associated with the RD design.

The main contribution of the paper is an investigation of how these statistical challenges manifest themselves in the body of published political science research using the RD design. We collect all articles using an RD design in the *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics* published from 2009 through 2018. We find that the distribution of published RD estimates exhibits some pathological features. Reported  $t$ -statistics bunch around, and especially just above, 1.96, corresponding to the conventional statistical significance level of five percent. Furthermore, estimated effect sizes are strongly associated with standard error sizes. This suggests selection pressure to obtain, report or publish results that are statistically significant at conventional levels.

A possible explanation for the pathological features lies with researcher discretion in analysis. Researcher discretion is an important concern because the RD design offers considerable leeway in how to estimate the effects, and researchers could leverage this leeway to search for significant findings, a practice sometimes called p-hacking. While the problems of p-hacking and publication bias are widely recognized, their importance for the RD design specifically has largely been overlooked; one recent exception is Gelman (2022). To investigate whether researcher discretion can explain the pathological features, we compare studies that use automated bandwidth selection procedures against those implementing non-automated procedures. Researcher discretion should be a greater concern with non-automated bandwidths, all else equal, as there are more choices to be made by researchers. However, we find *more* bunching for studies using automated bandwidth selection, suggesting that researcher discretion in bandwidth selection cannot readily explain our findings.

We next reanalyze all studies with available data in order to investigate the consequences of using inappropriate statistical procedures on the distribution of published results. Our reanalysis uses a standardized procedure based on current state of the art methods (Calonico, Cattaneo, & Titiunik, 2015). This aims to correct potential methodological shortcomings in studies using older methods and procedures, and to remove discretion in the analysis. The reanalysis does not meaningfully change the reported point estimates, but the estimated standard errors become larger on average, moving the  $t$ -statistics closer to zero. This

indicates that the body of published RD studies tends to systematically overestimate the statistical significance of their findings.

A contributing explanation for the pathological features could be selection pressure in the publication process. Researchers may abandon projects that fail to produce statistically significant results, or journal editors and referees may be reluctant to accept such results for publication. This type of publication bias would give rise to these pathological features even if researchers do not search through many specifications for statistically significant results. The problem has been well-documented in, for example, political science (A. S. Gerber & Malhotra, 2008a), experimental psychology (Open Science Collaboration, 2015), and economics (Brodeur, Cook, & Heyes, 2020).

The degree to which publication bias is consequential for the health of a literature depends on the statistical power of the analyses conducted in the field. In a body of underpowered studies, the share of rejected hypotheses that are false positives will typically be much greater than the nominal significance level. When a field consists primarily of underpowered studies and suffers from severe publication bias, a large majority of published findings could be false. Inappropriate inference procedures, such as biased standard error estimators, further inflate the number of false positives, exacerbating the problem. To investigate whether this is a concern for the body of RD studies, we conduct retrospective power analyses on all studies with available data. The exercise shows that most studies were not well powered to detect small- or moderate-sized effects. RD studies in political science indeed tend to be underpowered, sometimes severely so, making the concern over possible publication bias more alarming.

The findings in our paper are a stark contrast to the results of a similar study in economics by Brodeur et al. (2020). These authors found that the body of RD research in economics does not exhibit notable pathological features, especially compared to studies using instrumental variables. The difference between political science and economics might be puzzling given that applied empirical practice for RD estimation is similar in the two fields. However, as noted by Brodeur et al. (2020, p. 3646), differences between the fields have been documented before. For instance, A. S. Gerber and Malhotra (2008a) find that the ratio of tests just above the 1.96 threshold of statistical significance relative to those just below is 2 in political science, while Brodeur et al. (2020) find that the same ratio is only about 1.1 in economics. We investigate one potential explanation for the difference regarding RD research: namely that political scientists predominately use election data with the design, which often is a scarce source of data. Economists use register data or other microdata to a greater degree, which tend to be more plentiful, making the power issues less pronounced.

Taken together, our investigation suggests that many published findings using the RD design in political science may be exaggerated, if not spurious. Our analysis suggests that lack of sufficient statistical power combined with publication bias is a likely culprit. Power is

a particularly important concern for RD designs in political science as researchers predominantly use the design to study elections, where the amount of data is limited by the number of electoral districts and election cycles.<sup>1</sup> Although the RD design is an invaluable part of the methodological toolkit in political science, empirical researchers must nevertheless take the estimation challenges associated with the method seriously and properly address them. Our suggestions for empirical practice moving forward are that researchers should use appropriate, modern analysis procedures, that they should restrict focus to studies with sufficient power, and that they should pre-register analysis plans when possible.

## 2 Why is it hard to estimate RD effects?

The causal effect studied in the RD design is the difference between the expected treated and control potential outcomes at a cutoff where treatment assignment changes in a discontinuous fashion. If the conditional expectation function of the outcome were known, this effect can be calculated without error under an assumption that the conditional expectation functions of the potential outcomes are continuous at the cut point. In other words, the RD effect is *identified* under a continuity assumption. The assumption is reasonable in a wide variety of settings, such as when units cannot precisely control treatment assignment. This has earned the RD design its reputation as a method that produces credible findings. The concept of an RD design was introduced by Thistlethwaite and Campbell (1960). The formal identification results for the design were first derived by Hahn, Todd, and Klaauw (2001). An insightful and accessible discussion is provided by Cattaneo, Idrobo, and Titiunik (2020).

However, identification is not enough. The identification exercise presumed the conditional expectation function of the outcome was known, but it will generally not be. Learning about the causal effect requires estimating this function, and this estimation exercise may be unexpectedly challenging. While the identifying assumptions required by the RD design might be comparable in strength to those required by randomized controlled trials (RCTs), the *statistical* challenges under the RD design are better compared with those we typically face in other types of observational studies. The main challenge is that we are interested in the value of the potential outcome functions when evaluated at the cutoff, but there will generally be no observations exactly at the cutoff. Therefore, researchers need to rely on observations away from the cutoff for estimation, and these observations may not be informative of the quantities of interest.

The standard way to proceed is to assume that the potential outcome functions are smooth, meaning that they do not change too quickly. This implies that observations close to the cutoff are somewhat informative of the value of the functions at the cutoff, so they

---

<sup>1</sup>We thank Tara Slough for suggesting this point to us.

can be used for estimation. Researchers then face a trade-off. If we estimate the RD effect using only observations very close the cutoff, we ensure that the observations are relevant, but doing so discards most of the data, rendering the estimates less precise, as measured by for example variance. By contrast, if we include observations farther away from the cutoff, precision improves, but the relevance of the observations decreases. We can interpret this as a bias–variance trade-off, where the bias is governed by the relevance of the observations and the variance by their numbers.

No matter how one resolves this trade-off, the consequence is that the nominal sample size (i.e., the number of rows in the data set) is not a good measure of the effective sample size (i.e., the amount of useful information in the data set). Even if identification in an RD design might be almost as credible as in a randomized experiment, we would often need a sample size that is orders of magnitude larger to estimate the causal effect in an RD design to the same level of accuracy as in an experiment. The bias–variance trade-off also introduces researcher discretion in that there is no intrinsically correct way to resolve the trade-off, and researchers can, purposefully or inadvertently, use this leeway to search for significant findings.

The bias–variance trade-off is inherent to the RD design, and therefore inescapable, but several methods have been developed to help researchers navigate it. The most prevalent approach is to carefully balance bias and variance with the aim to maximize accuracy, as measured by mean square estimation error. The bandwidth selection method described by Imbens and Kalyanaraman (2011) is an early such example. While this and other similar methods do improve accuracy, they cannot escape the inherent limitations of the data; the effective sample size often remains small even when the nominal sample size is large. Besides improving accuracy, these methods also reduce researcher discretion, possibly limiting researchers’ ability to do specification searching. Of course, researcher discretion is not entirely removed. Researchers may, for example, still be able to choose between different bandwidth selection methods, and decide on other aspects of the analysis.

The use of bandwidth selection methods to maximize accuracy provides a partial solution to some of the most pressing problems associated with RD designs, but in doing so, they create a new one. To conduct hypothesis tests or construct confidence intervals, researchers need to estimate the accuracy of the RD point estimate. Conventional methods used to gauge this accuracy rely on an assumption that the bias of the estimator is negligible compared to its variance, as would be the case in, for example, a randomized experiment. However, when a bandwidth selection method balances bias and variance to maximize accuracy, the bias is not negligible, and it must therefore be accounted for when drawing inference. Researchers sometimes neglect doing so, with the consequence that their hypothesis tests and confidence intervals are misleading even when the sample is large.

There are predominately three approaches to address this concern. The first approach is so-called undersmoothing, meaning that one uses a bandwidth that is smaller than the

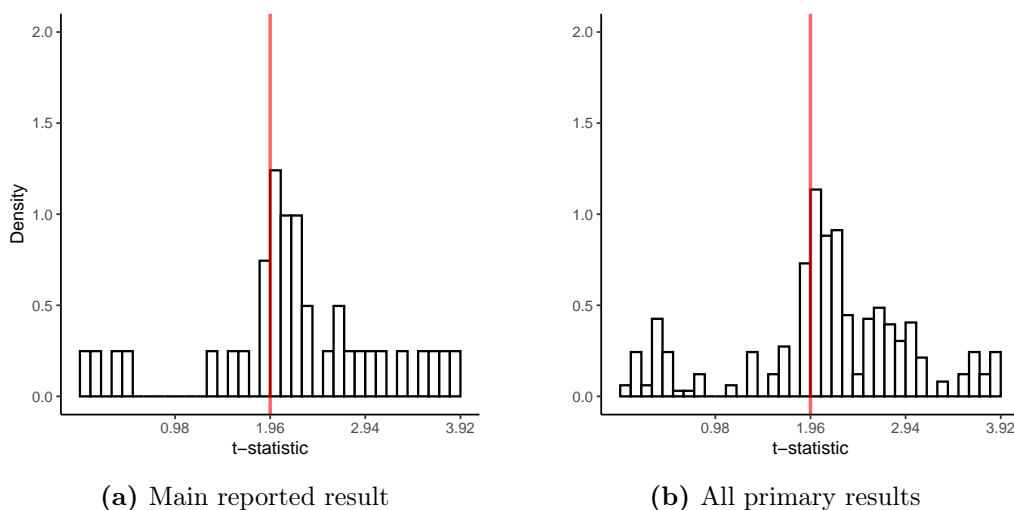


optimal one. If done correctly, this will ensure that the bias indeed is negligible in large samples, so researchers can use conventional methods to construct hypothesis tests and confidence intervals. The obvious downside is that the accuracy of the point estimator is worse compared to when using the optimal bandwidth.

The second approach is to adjust the inference procedure for the bias. One can show that the bias when using the optimal bandwidth can be estimated using a similar approach as for the point estimator. This bias estimate can be used to adjust the point estimate before the construction of confidence intervals and hypothesis tests, allowing researchers to address the bias and still use the optimal bandwidth. To ensure that the bias adjustment itself does not affect the overall precision, one must ensure that the bias is estimated with considerably higher precision than the precision of the point estimator. This cannot be ensured in finite samples, but it can be ensured asymptotically if the bandwidth used for the bias adjustment is considerably larger than the bandwidth used for the point estimator. However, Calonico, Cattaneo, and Titiunik (2014) point out that it might be undesirable to use a larger bandwidth for the bias adjustment, and that the finite sample behavior might nevertheless be poor. The alternative approach described by Calonico et al. (2014), which is becoming the standard approach used in political science, is to also estimate the precision of the bias adjustment and incorporate this into the variance estimator. This allows researchers to use bias adjustments that have similar precision as the point estimator (e.g., as would be the case when the same bandwidth is used for both point and bias estimation), and it should improve finite sample performance.

The third approach is to adjust the inference procedure with the worse-case bias consistent with the observed data. The worst-case is taken with respect to a stipulated class of functions, in which the true conditional expectation function is assumed to exist. This approach obviates the need to estimate the bias, which means that we have to rely on fewer asymptotic approximations and assumptions than with the second approach. As a consequence, the third approach is expected to perform better in terms of coverage in finite samples, at the cost of typically being more conservative. The method described by Armstrong and Kolesár (2020) is an example of this approach.

Taken together, the estimation challenges associated with the RD design are immense, comparable to those faced in other types of observational studies. Yet, being attracted by its weak identification assumptions, researchers may overlook these challenges. The concerns become even more pressing when the nominal sample sizes are small, as they often are for RD studies in political science. In particular, the most common type of RD study in political science uses election cutoffs to study the effect of incumbency or other characteristics of candidates or parties. The sample size is here limited by the number of election districts and the frequency of the elections, which both tend to be small. For example, if we are interested in contemporary American gubernatorial elections, we are limited to around 200 observations, of which most will not be close elections and thus not



**Figure 1:** Distribution of  $t$ -statistics among published RD studies in political science

relevant when estimating the RD effect. These estimation challenges and the relatively limited number of observations in election RDs, in particular, motivate us to take a closer look at the state of RD research in political science.

### 3 The State of RD Studies in Political Science

We collected all studies using an RD design published in the *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics* between 2009 and 2018. We included studies implementing an RD design as the primary empirical strategy in an applied setting, in addition to studies where an RD design complements another design (e.g., Broockman & Ryan, 2016). We excluded studies primarily making methodological contributions (e.g., Cattaneo, Keele, Titiunik, & Vazquez-Bare, 2016). There were 45 studies in total that satisfied these inclusion criteria. Section S1 in the supplement describes the sample selection strategy in more detail and contains a list of all 45 studies.

Figure 1 presents the distribution of  $t$ -statistics for the published findings among all 45 studies in our sample. The first panel presents the distribution of the main reported result of each article, meaning that each article contributes one  $t$ -statistic. The second panel presents the distribution of all results that were referenced in the articles' abstracts, which we take as a proxy for being a primary result of an article. There are 80  $t$ -statistics in total in the second panel, meaning that each article contributes 1.78 statistics on average. To avoid overrepresentation of articles that present many results, the  $t$ -statistics in the second panel are weighted by the inverse of the total number of results in each study. For the majority of the studies, we either use reported  $t$ -statistics or reconstruct  $t$ -statistics

using the ratio of coefficients and standard errors. For the small number of the studies where neither  $t$ -statistics nor standard errors were reported, we reconstruct  $t$ -statistics using reported  $p$ -values, confidence intervals and point estimates, or using replication code and data provided by the authors.

The histograms in both panels of Figure 1 demonstrate clustering around the value 1.96. This corresponds to a significance level of 5%, which is the conventional threshold for calling a result “significant.” Particularly noteworthy is the substantial imbalance in density to the right of 1.96, suggesting that results that clear the 5% significance level threshold are artificially favored. The density of  $t$ -statistics increases somewhat before the 1.96 threshold, which could indicate that almost significant results are also favored. Nevertheless, the spike in density to the right of 1.96 is pathological, in the sense that we would not expect to observe this pattern of  $t$ -statistics occurring naturally.

The number of RD studies in our population is relatively small, and it is perhaps conceivable that this distribution of  $t$ -statistics could arise with somewhat high probability even when there are no real pathologies in the literature. To assess this in a more formal way, we turn to the so-called caliper test described by A. S. Gerber and Malhotra (2008a, 2008b), which was also used by Brodeur et al. (2020).<sup>2</sup> The test compares the number of studies just above and just below the 1.96 threshold that corresponds to statistical significance at the 5% level. Under the assumption (null hypothesis) that the 5% significance level is irrelevant for researchers’ and editors’ publication decisions, the 1.96 threshold should be immaterial, and we would expect approximately the same number of studies just above and just below it. However, the 1.96 threshold becomes relevant for the publication decision if there is  $p$ -hacking or publication bias. In this case, we would not be surprised to see more studies just above the threshold. Following A. S. Gerber and Malhotra (2008a, 2008b), we include studies within windows around the cutoff with radii (i.e., half-widths or calipers) of 10%, 15% and 20% of the threshold itself, here corresponding to 0.196, 0.294 and 0.392, respectively. A smaller window is better targeted towards the 1.96 threshold, but has lower power. To account for the likely dependence between estimates from the same study, we include only the main estimate from each study in this test.

Table 1 presents the results from the caliper tests. For all three window sizes, there are approximately three times as many studies just above the threshold than just below it. We calculate the probability of observing an imbalance at least this extreme under the null hypothesis of no real pathologies in the RD literature by using a binomial distribution with equal success probability, capturing that we would expect a study to be equally likely to be just above and below the threshold under the null. Following A. S. Gerber and Malhotra (2008a, 2008b), the test is one-tailed, reflecting the fact that we feel comfortable ruling out, before seeing the data, that researchers or editors actively favor insignificant findings. The  $p$ -values of the caliper tests range between 3% and 7%, depending on the window

---

<sup>2</sup>We thank one of the reviewers and the editor for the suggestion of adding these caliper tests.

size, indicating that it is unlikely that we would observe this large of an imbalance in the  $t$ -statistics if there were no real pathologies in the RD literature. These findings are in line with A. S. Gerber and Malhotra (2008a, 2008b). However, as noted in the introduction, the findings are in contrast to the findings of Brodeur et al. (2020), who document fewer pathological features for the RD design in economics than other commonly used methods.

**Table 1:** Caliper tests of imbalance of  $t$ -statistic among RD studies

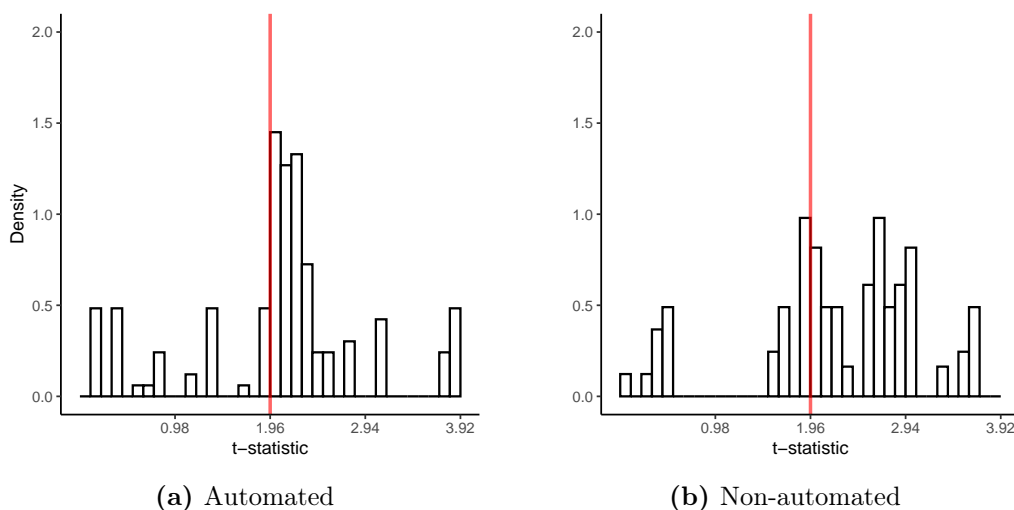
Caliper	Below	Above	$p$ -value
10%	3	9	0.073
15%	4	11	0.059
20%	5	14	0.032

## 4 Researcher discretion

A possible explanation for the bunching of  $t$ -statistics around the value 1.96 observed in the previous section is researcher discretion. Such discretion could be used to seek statistically significant results by searching through specifications and analysis procedures. When done deliberately, the practice is called p-hacking, but researchers can conduct specification searching inadvertently. No matter whether it is deliberate or not, searching for significant results invalidates hypothesis tests,  $p$ -values and confidence intervals.

It is difficult to directly gauge the influence of researcher discretion and possible p-hacking. We do not observe all specifications researchers tried before finding the one reported in their published article, and unusual specification choices can often be rationalized after the fact. In an effort to circumvent this problem, we take advantage of the fact that different methods of bandwidth selection provide different levels of researcher discretion. An automated bandwidth selection procedure, such as those described by Imbens and Kalyanaraman (2011) and Calonico et al. (2014), leaves less room for specification searching. This is in contrast to non-automated procedures and global polynomial specifications, which involve many specification choices. Table S3 in the supplement describes the bandwidth selection procedures used by the studies in our investigation. This table shows that while automated procedures have seen increased use over time, roughly half of all studies throughout the studied period use non-automated procedures.

Figure 2 presents histograms of  $t$ -statistics disaggregated by studies that use and do not use an automated bandwidth selection procedure. The first panel consists of studies using automated methods, and we see a clear spike to the right of 1.96, similar to the finding in the previous section. The second panel, which consists of studies using non-automated methods, does not present a clear spike, although there are more studies to the right than



**Figure 2:** Distributions of  $t$ -statistics disaggregated by bandwidth selection procedure

to the left of the cutoff. To the degree that the bandwidth selection method is a good proxy of researcher discretion, this finding suggests that discretion is not the driving force of the pathological features we observed in the previous section.

We conduct the caliper tests described in the previous section in the two groups of studies using automated and non-automated bandwidth selection procedures. The results are presented in Section S5 of the supplement. In summary, there are almost the same number of  $t$ -statistics just above as below the 1.96 threshold within all windows for the studies using non-automated procedures, with  $p$ -values around 50%, indicating that this is not an unusual observation under the null. However, there is only one  $t$ -statistic just below the threshold for studies using automated procedures, irrespectively of window size, and there are 6 to 10  $t$ -statistics just above the threshold, depending on window size. The  $p$ -values are between 1% and 6%, indicating that this would be an unusual observation under the null. This corroborates the insights given by the histograms. However, note that the samples are smaller in these disaggregate tests, so power is reduced.

These results should be interpreted with caution. When researchers use non-automated bandwidth selection methods, they are generally expected to report the estimates using several different bandwidths, with the expectation that the estimates are similar for most of those bandwidths. This limits the scope of specification searching even when using a non-automated method, making the bandwidth selection method less useful as a proxy for researcher discretion. There are also other sources of researcher discretion. For example, unless researchers have pre-committed to a bandwidth selection method before having access to the data, the decision to use a particular automated bandwidth selection method may itself be part of the specification search. Researchers also have discretion over which

questions to investigate, and they may search through a large set of different outcomes, trying to find one that is statistically significant at conventional levels. While the current investigation does not allow us to rule out all types of researcher discretion as an explanation, it suggests that specification searching using bandwidth selection is not a major driving force of the pathological features we saw in the previous section.

## 5 Reanalysis

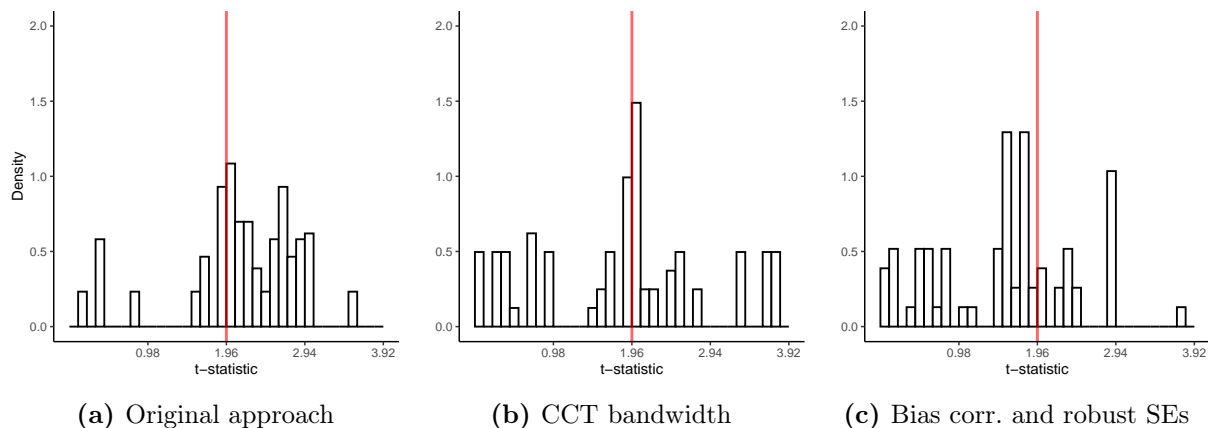
We next seek to understand what these studies, all else equal, might have looked like had they all been analyzed with the same, modern approach. This serves two purposes. First, it provides additional insights about whether researcher discretion is a contributing factor to the pathological features. By reanalyzing all studies using a standardized procedure, we remove many sources of researcher discretion. Second, some studies used procedures that do not properly address the statistical concerns outlined in Section 2, potentially providing misleading results. A reanalysis addresses many, if not all, of these statistical concerns.

### 5.1 Methodology

Replication data are required to conduct this reanalysis. We were able to obtain such data for 36 out of the 45 studies in our sample. Section S2 in the supplement describes how we collected the replication data and lists the studies for which we failed to obtain such data, including brief descriptions of why we failed to do so. The most common reason was the use of proprietary data, which the original authors were prohibited from sharing.

Our primary reanalysis uses the default settings in the R package `rdrobust` (Calonico et al., 2015). The procedure implemented in this package has well-understood theoretical properties and attempts to address the concerns discussed in Section 2. The reanalysis results we present in the main body of the paper exclude all studies that originally used the `rdrobust` package to estimate the effects, as a re-analysis would bring few new insights for these studies. This results in 39 estimates across 25 studies. Thus, the investigation presented in this section focuses on these 25 studies. The re-analysis for the full set of 36 studies with replication data is presented in Section S8 in the supplement.

One limitation of the approach by Calonico et al. (2015) is that it is motivated by an asymptotic approximation. This means that confidence intervals and hypothesis tests may not be valid in finite samples. For example, Eckles, Ignatiadis, Wager, and Wu (2020) reports Monte Carlo simulations suggesting that even when  $n = 1,000$ , coverage for nominal 95% confidence intervals can be less than 92%. For this reason, we also conduct reanalyses using the `rdhonest` procedure by Armstrong and Kolesár (2020), for which asymptotic approximations feature less heavily, as discussed in Section 2. However, `rdhonest` also has limitations. Unlike `rdrobust` which adapts to the degree of smoothness in the re-



**Figure 3:** Distributions of  $t$ -statistics among replicated studies by method of analysis

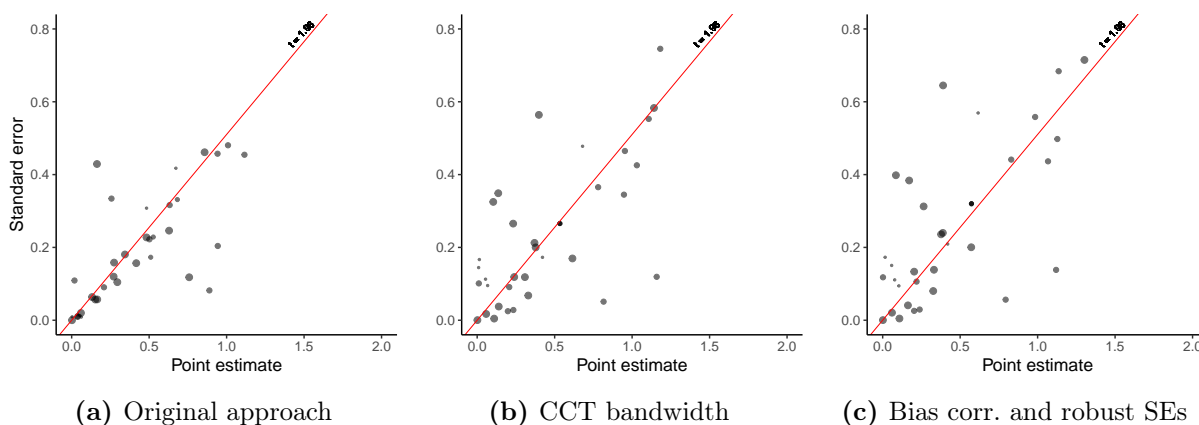
gression function, `rdhonest` requires that the researcher specifies, or estimates, a lower bound on smoothness. In our setting, we use the global quartic regression implemented in `rdhonest` to estimate a bound on smoothness, which requires stronger assumptions than if the smoothness bound was selected based on prior knowledge.

## 5.2 Results

Figure 3 presents the distribution of  $t$ -statistics among the 39 estimates from the 25 studies included in the main replication exercise. Panel 3a replicates Figure 1b for the 39 estimates, containing the reported  $t$ -statistics using the analysis procedures used in the original studies.

Panel 3b re-analyzes all 39 estimates using the default bandwidth selection method implemented in `rdrobust` (“`mserd`”) but does not make any adjustments for bias for the point estimator nor any adjustments for the standard errors. For this reason, these  $t$ -statistics will be too large as the accuracy of estimates is overestimated, potentially making them misleading. However, if specification searching in bandwidth selection was an important driver, there would be noticeable differences between the first and second panels. The figure demonstrates that the distribution shifts to the left compared to the first panel, but a clear spike remains at the value 1.96 in the second panel. This provides additional support to the conclusion that specification searching over bandwidths is not the main driver of the findings.

Panel 3c re-analyzes the studies using the same bandwidth selection method as in the second panel but also implements the bias adjustment for the point estimator and adjusts the estimated standard errors to account for the added imprecision of the bias adjustment. Here, we see a clear shift to the left, indicating that these studies have systematically overestimated the accuracy of their estimates.



**Figure 4:** Funnel plots for replicated studies by method of analysis

We can disaggregate the information in Figure 3 using a funnel plot, which is a scatter plot with standardized point estimates and standardized estimated standard errors on the two axes. Figure 4 presents such a funnel plot for the 39 estimates in the replication study.<sup>3</sup> We use the sample variance of the outcome among control units for the standardization. The boundary for statistical significance at the 5% level is a line emanating from the origin with a slope of 1.96, here drawn in red, where estimates to the right of this line attain statistical significance at the conventional level. Like in Figure 3, the first panel presents the estimates from the original studies, the second panel uses the default bandwidth selection method in `rdrobust` but no bias adjustment, and the third panel implements the bias adjustment.

We find that the re-analysis does not systematically change point estimates, but it does have meaningful consequences for the estimated standard errors. The standard errors are systematically larger in Figure 4b than in Figure 4a. This is partly because some studies use global polynomial specifications to estimate the RD effect. A global polynomial specification does not impose any bandwidth restriction on the sample, making the bandwidth selected by `rdrobust` dramatically smaller, leading to larger standard errors. The standard errors are even larger in Figure 4c, reflecting the added imprecision introduced by the bias adjustment.

We note that the estimates cluster around the red line corresponding to a  $t$ -statistic of 1.96, especially in the first panel. Echoing the logic of A. S. Gerber, Green, and Nicker-son (2001), it is hard to rationalize such a strong positive correlation between the point estimates and standard errors. In theory, one explanation could be that researchers are tremendously adept at power analysis, so they are able to target the sample size to have just

<sup>3</sup>Because the funnel plots require standardization, we cannot construct these plots for studies without replication data, which is why they were not used in Section 3.



sufficient power to detect the effect of interest. However, this explanation appears improbable; there are many uncertainties involved with a power analysis, and researchers rarely have precise control over the sample size in an RD study. A more probable explanation is selection pressure on what type of results get reported and published.

The results of the reanalysis are largely unchanged when we use the `rdhonest` procedure by Armstrong and Kolesár (2020) in place of `rdrobust`. Section S8 in the supplement reports the results using this alternative approach. The version of the `rdhonest` package we used for the re-analysis cannot account for statistical clustering, so our replication using this package is restricted to studies that did not originally use clustered standard errors. Because the `rdhonest` package uses a worst-case bias adjustment, conventional  $t$ -statistics will not capture the adjustment made here. To make the results from `rdhonest` comparable to the other results in this paper, we derive pseudo  $t$ -statistics based on the width of the confidence intervals produced by the `rdhonest` package. Substantively, the distribution of reanalysis results is similar to those produced by `rdrobust`: the pseudo  $t$ -statistics shift to the left, and the pseudo standard errors systematically increase.

## 6 Power Analysis

Having well-powered studies is one of the best defenses against the concerns highlighted in this paper. Selection pressure for significant findings in a body of poorly powered studies could make most of the reported results spurious. This is because the probability of rejecting a false null hypothesis is not much greater than rejecting a true null hypothesis when power is low, meaning that the proportion of false positives of all positive findings in the published literature will be approximately the same as the proportion of true null hypothesis in the body of conducted studies. In some literatures, most investigated (alternative) hypotheses are false, meaning that most published positive findings also will be false if the power is low. But in a body of well-powered studies, the true positives will be a disproportionately large share of the overall positive findings, decreasing the proportion of published false positives. Furthermore, a well-powered study can in some cases make researcher discretion less consequential, because there is less variability in the estimate to exploit. Therefore, as noted in Section 2, it is worrying that the RD design is unusually demanding with respect to sample size, and even samples that on the surface appear large can be poorly powered.

To investigate the extent to which this is a relevant concern among RD studies in political science, we conducted retrospective power analyses for all 36 studies with replication data, comprising 64 separate analyses. Using the power analysis method implemented in the R package `rdpower` by Cattaneo, Titiunik, and Vazquez-Bare (2019), we estimate the power of a two-tailed test at the 5% significance level based on a central limit approximation for the sampling distribution. The power analyses presume that  $p$ -values will be constructed with the bias adjustment and robust standard errors as implemented by `rdrobust`. As this

**Table 2:** Proportion of analyses achieving 60%, 80% and 95% power by effect size

Effect size	Power		
	60%	80%	95%
0.1	0.17	0.15	0.10
0.2	0.31	0.22	0.18
0.5	0.65	0.56	0.48
0.8	0.78	0.70	0.65

procedure is only asymptotically valid, our power calculations may be optimistic relative to approaches that control the false positive rate in finite samples.

We investigate power with respect to four difference effect sizes, ranging from small to large effects. We measure effect size by Cohen’s  $d$ , which is the treatment effect standardized by the standard deviation of the outcome (Cohen, 1988). We use the standard deviation of the outcome of control units within the default bandwidth in the `rdrobust` package for this standardization. We investigate the effect sizes 0.1, 0.2, 0.5 and 0.8. While 0.5 is conventionally labelled as a medium-sized effect, modern social science tends to study effects of smaller sizes. For example, the What Works Clearinghouse, which is a governmental program that collects and reviews evidence of the effectiveness of various policies, labels an effect size of 0.25 as “substantively important” in their handbook (What Works Clearinghouse, 2017).

Table 2 presents the results from the power analyses. The cells give the proportion of the 64 analyses that attain the power specified by the columns for the effect size specified by the rows. For example, we see that only 17% of the analyses attain 60% power to detect a standardized effect of 0.1. We weigh the analyses in the same way as in Figure 1b to account for the fact that some studies conducted more analyses than others.

The table shows that these studies are overall poorly powered to detect anything but large effects. Only 22% of the analyses achieve 80% power to detect a 0.2 effect size. Recall that 80% power is the conventional level that researchers often strive to achieve when designing a study. Hence, about four out of five RD studies are not properly powered for effect sizes that in many applications would be considered noteworthy. For the effect sizes of 0.5 and 0.8, the shares of properly powered studies increase to 56% and 70%, respectively. Both 0.5 and 0.8 would be very large effects in most modern literatures in political science, so the fact that almost half of the studies are not well-positioned to detect such effects gives an indication of the severity of the problem. Overall, the body of RD studies is alarmingly underpowered.

As previously discussed, one possible reason for the lack of power in RD studies in political science lies with the fact that many RD studies in the field use data from elections. The number of elections in a given region is necessarily finite, often with no way to accumulate

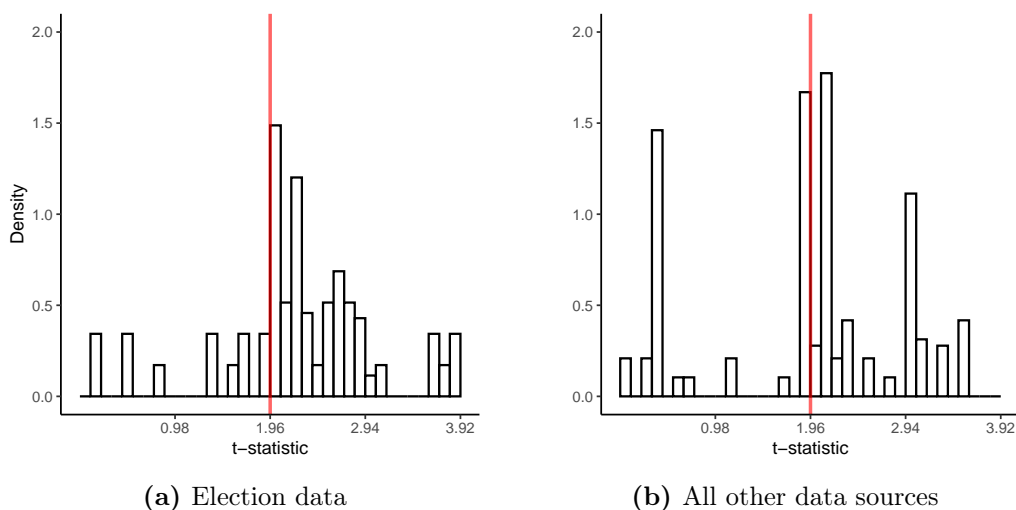
**Table 3:** Proportion of estimates achieving 60%, 80% and 95% power by effect size

(a) Studies Using Elections				(b) Studies Not Using Elections			
Effect size	Power			Effect size	Power		
	60%	80%	95%		60%	80%	95%
0.1	0.08	0.08	0.04	0.1	0.40	0.35	0.28
0.2	0.21	0.13	0.10	0.2	0.55	0.42	0.40
0.5	0.59	0.46	0.40	0.5	0.80	0.80	0.69
0.8	0.78	0.66	0.59	0.8	0.80	0.80	0.80

more data of a similar type through additional data collection efforts. We investigate this as a potential explanation by examining the power of such studies. Of the studies in our investigation, 68.9% use electoral data, and among studies with replication data, 72.2% use electoral data. Table 3 presents power for the studies disaggregated by data source. We see that studies in political science that use electoral data have considerably lower power than those that do not. Along with general differences in disciplinary norms in replicability across fields, this helps to explain why our results are more pessimistic for the RD design than those of Brodeur et al. (2020) in the field of economics.

To get a sense of how consequential these power issues are for the pathological features we observed in Section 3, we consider the distributions of  $t$ -statistics separately for studies using election data and those that do not, in a similar fashion as the investigation in Section 4. The distributions are presented in Figure 5. We see that there are noticeably more  $t$ -statistics just above the 1.96 threshold for studies using election data, while the  $t$ -statistics are more balanced around the threshold for the other studies. Although, there is a noticeable spike at the threshold for the second set of studies as well. This observation is consistent with insufficient statistical power being the main driver of the pathological features, but this investigation is far from conclusive. It could, for example, be that election research is more competitive in political science, meaning that researchers have a greater incentive to p-hack and that editors are more selective. Additionally, we are limited in the conclusions we draw about studies using non-election data given the limited number of published papers using such data.

While power is a particularly important consideration for the RD design, given its hunger for data, the issues we document here are not unique to RD studies. Arel-Bundock, Briggs, Doucouliagos, Aviña, and Stanley (2022) use meta-analytic evidence to estimate that approximately 1 in 10 tests reported in political science have at least 80% power. This analysis relies on the assumption that studies in a given literature share a common mean effect (inclusive of bias), which may understate the power of many studies in the presence of effect heterogeneity or differential bias across studies. Nevertheless, the striking degree of the documented issues suggest that power issues are prevalent throughout the field.



**Figure 5:** Distributions of  $t$ -statistics disaggregated by data source

## 7 Concluding remarks

The body of published political science research using the RD design exhibits pathological features consistent with selective reporting and publishing based on whether results clear artificial significance thresholds. More than half of the studies in our sample do not properly estimate the accuracy of their estimates, leading to the associated hypothesis tests and confidence intervals being misleading. Most of the studies are also underpowered, and are able to detect only large effects. Taken together, this paints a somber picture of the state of applied studies using the RD design. Our results suggest that many published findings using the design are exaggerated if not altogether spurious.

The conclusion is that researchers using the RD design must take the estimation challenges associated with the design more seriously. They should make sure to conduct hypothesis tests and construct confidence intervals in a well-motivated way that properly reflects the accuracy of the finding. The procedures described by Calonico et al. (2014) and Armstrong and Kolesár (2020) represent significant progress toward this goal.

Furthermore, statistical power will often be a first-order concern with the RD design, and researchers should make sure that they have sufficiently large samples to have a good chance to detect effects of sizes relevant to the question at hand. When gauging power, researchers should remember that the nominal sample size is not relevant in an RD study, and they should instead consider how much information there is about the potential outcomes close to the cutoff. Sample size is often beyond the control of the researcher in RD studies; if the accessible sample is too small, researchers should ask whether it is appropriate to conduct the study at all. The decision to abandon a study due to concerns about power

should be taken before running the analysis and observing the estimated effect.

Covariates can sometimes be used to improve precision in an RD study without collecting additional observations, which could alleviate some of the concerns about lack of power with the RD design (Calonico, Cattaneo, Farrell, & Titiunik, 2019). However, covariate adjustment involves many largely arbitrary specification choices, which would make it easier to p-hack. It is unclear if the body of RD studies overall would benefit from making covariate adjustment an expected part of the analysis with the RD design. Indeed, one of the central benefits of the RD design is exactly that identification does not require background information about the units, other than the running variable that decides treatment assignment.

Journals should consider taking power into account when making publication decisions. If a study is severely underpowered, it may not contribute much to the literature even if it has nominally statistically significant results. Similarly, a well-powered study often provides useful insights no matter if its results are statistically significant at conventional levels, because confidence intervals will typically be narrow. Being more mindful about power would alleviate publication bias, making the body of published results more reliable and informative.

Another way to alleviate some of these concerns is preregistration for studies using the RD design. Given that most RD studies are retrospective, it might be difficult for researchers to credibly show that all analyses were registered before they were conducted. However, insofar as it is possible, having a full record of analyses that were intended would be helpful for the accumulation of knowledge. They reveal the absence of findings from the published literature due to the “file drawer” problem that disproportionately afflicts insignificant findings, either through the filter of publication or through specification searching. Scholars evaluating a literature (e.g., in meta-analysis) could use these preregistrations to help to undo distortions introduced through the various filters associated with the production and dissemination of knowledge. Preregistration is also useful to address inadvertent p-hacking, as researchers are less likely to accidentally or thoughtlessly deviate from their original plans when they are recorded.

## References

- Anzia, S. F., & Berry, C. R. (2011). The jackie (and jill) robinson effect: Why do congresswomen outperform congressmen? *American Journal of Political Science*, *55*(3), 478–493.
- Arel-Bundock, V., Briggs, R. C., Doucouliagos, H., Aviña, M. M., & Stanley, T. (2022). *Quantitative political science research is greatly underpowered*. (OSF 7vy2f)
- Ariga, K. (2015). Incumbency disadvantage under electoral rules with intraparty competition: Evidence from japan. *The Journal of Politics*, *77*(3), 874–887.

- Armstrong, T., & Kolesár, M. (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics*, 11(1), 1–39.
- Boas, T. C., & Hidalgo, F. D. (2011). Controlling the airwaves: Incumbency advantage and community radio in Brazil. *American Journal of Political Science*, 55(4), 869–885.
- Boas, T. C., Hidalgo, F. D., & Richardson, N. P. (2014). The spoils of victory: Campaign donations and government contracts in Brazil. *The Journal of Politics*, 76(2), 415–429.
- Bohlken, A. T. (2018). Targeting ordinary voters or political elites? why pork is distributed along partisan lines in India. *American Journal of Political Science*, 62(4), 796–812.
- Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11), 3634–3660.
- Brollo, F., & Nannicini, T. (2012). Tying your enemy’s hands in close races: The politics of federal transfers in Brazil. *American Political Science Review*, 106(4), 742–761.
- Broockman, D. E., & Ryan, T. J. (2016). Preaching to the choir: Americans prefer communicating to copartisan elected officials. *American Journal of Political Science*, 60(4), 1093–1107.
- Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2019). Regression discontinuity designs using covariates. *The Review of Economics and Statistics*, 101(3), 442–451.
- Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6), 2295–2326.
- Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015). rdrobust: An R package for robust nonparametric inference in regression-discontinuity designs. *R Journal*, 7(1), 38–51.
- Carson, J. L., & Sievert, J. (2017). Congressional candidates in the era of party ballots. *The Journal of Politics*, 79(2), 534–545.
- Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2020). *A practical introduction to regression discontinuity designs* (Vol. 1). Cambridge University Press.
- Cattaneo, M. D., Keele, L., Titiunik, R., & Vazquez-Bare, G. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, 78(4), 1229–1248.
- Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2019). Power calculations for regression-discontinuity designs. *The Stata Journal*, 19(1), 210–245.
- Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2020). The regression discontinuity design. In L. Curini & R. Franzese (Eds.), *The SAGE handbook of research methods in political science and international relations* (Vol. 2, pp. 835–857). London: SAGE Publications.
- Caughey, D., Dafoe, A., & Seawright, J. (2017). Nonparametric combination (npc): A framework for testing elaborate theories. *The Journal of Politics*, 79(2), 688–701.
- Caughey, D., Warshaw, C., & Xu, Y. (2017). Incremental democracy: The policy effects of partisan control of state government. *The Journal of Politics*, 79(4), 1342–1358.
- Cavaille, C., & Marshall, J. (2019). Education and anti-immigration attitudes: Evidence from compulsory schooling reforms across Western Europe. *American Political Science Review*, 113(1), 254–263.
- Clinton, J. D., & Sances, M. W. (2018). The politics of policy: The initial mass political effects of Medicaid expansion in the states. *American Political Science Review*, 112(1), 167–185.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.
- Coppock, A., & Green, D. P. (2016). Is voting habit forming? new evidence from experiments and regression discontinuities. *American Journal of Political Science*, 60(4), 1044–1062.

- Croke, K., Grossman, G., Larreguy, H. A., & Marshall, J. (2016). Deliberate disengagement: How education can decrease political participation in electoral authoritarian regimes. *American Political Science Review*, *110*(3), 579-600.
- Dahlgaard, J. O. (2018). Trickle-up political socialization: The causal effect on turnout of parenting a newly enfranchised voter. *American Political Science Review*, *112*(3), 698-705.
- de Benedictis-Kessner, J. (2018). Off-cycle and out of office: Election timing and the incumbency advantage. *The Journal of Politics*, *80*(1), 119-132.
- de Benedictis-Kessner, J., & Warshaw, C. (2016). Mayoral partisanship and municipal fiscal policy. *The Journal of Politics*, *78*(4), 1124-1138.
- de Kadt, D. (2017). Voting then, voting now: The long-term consequences of participation in south africa's first democratic election. *The Journal of Politics*, *79*(2), 670-687.
- Dunning, T., & Nilekani, J. (2013). Ethnic quotas and political mobilization: Caste, parties, and distribution in indian village councils. *American Political Science Review*, *107*(1), 35-56.
- Eckles, D., Ignatiadis, N., Wager, S., & Wu, H. (2020). *Noise-induced randomization in regression discontinuity designs*. (arXiv:2004.09458)
- Eggers, A. C. (2017). Quality-based explanations of incumbency effects. *The Journal of Politics*, *79*(4), 1315-1328.
- Eggers, A. C., Fowler, A., Hainmueller, J., Hall, A. B., & Snyder Jr., J. M. (2015). On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races. *American Journal of Political Science*, *59*(1), 259-274.
- Eggers, A. C., Freier, R., Grembi, V., & Nannicini, T. (2018). Regression discontinuity designs based on population thresholds: Pitfalls and solutions. *American Journal of Political Science*, *62*(1), 210-229.
- Eggers, A. C., & Hainmueller, J. (2009). Mps for sale? returns to office in postwar british politics. *American Political Science Review*, *103*(4), 513-533.
- Eggers, A. C., & Spirling, A. (2017). Incumbency effects and the strength of party preferences: Evidence from multiparty elections in the united kingdom. *The Journal of Politics*, *79*(3), 903-920.
- Erikson, R. S., Folke, O., & Snyder, J. M. (2015). A gubernatorial helping hand? how governors affect presidential elections. *The Journal of Politics*, *77*(2), 491-504.
- Ferwerda, J., & Miller, N. L. (2014). Political devolution and resistance to foreign rule: A natural experiment. *American Political Science Review*, *108*(3), 642-660.
- Fiva, J. H., & Smith, D. M. (2018). Political dynasties and the incumbency advantage in party-centered environments. *American Political Science Review*, *112*(3), 706-712.
- Folke, O., Hirano, S., & Snyder, J. M. (2011). Patronage and elections in u.s. states. *American Political Science Review*, *105*(3), 567-585.
- Folke, O., Persson, T., & Rickne, J. (2016). The primary effect: Preference votes and political promotions. *American Political Science Review*, *110*(3), 559-578.
- Folke, O., & Snyder, J. M. (2012). Gubernatorial midterm slumps. *American Journal of Political Science*, *56*(4), 931-948.
- Fourinaies, A., & Hall, A. B. (2014). The financial incumbency advantage: Causes and consequences. *The Journal of Politics*, *76*(3), 711-724.
- Friedman, J. N., & Holden, R. T. (2009). The rising incumbent reelection rate: What's gerrymandering got to do with it? *The Journal of Politics*, *71*(2), 593-611.

- Galasso, V., & Nannicini, T. (2011). Competing on good politicians. *American Political Science Review*, 105(1), 79-99.
- Gelman, A. (2022). Criticism as asynchronous collaboration: An example from social science research. *Stat*, 11(1), e464.
- Gerber, A. S., Green, D. P., & Nickerson, D. (2001). Testing for publication bias in political science. *Political Analysis*, 9(4), 385-392.
- Gerber, A. S., & Huber, G. A. (2010). Partisanship, political control, and economic assessments. *American Journal of Political Science*, 54(1), 153-173.
- Gerber, A. S., Kessler, D. P., & Meredith, M. (2011). The persuasive effects of direct mail: A regression discontinuity based approach. *The Journal of Politics*, 73(1), 140-155.
- Gerber, A. S., & Malhotra, N. (2008a). Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science*, 3(3), 313-326.
- Gerber, A. S., & Malhotra, N. (2008b). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*, 37(1), 3-30.
- Gerber, E. R., & Hopkins, D. J. (2011). When mayors matter: Estimating the impact of mayoral partisanship on city policy. *American Journal of Political Science*, 55(2), 326-339.
- Gulzar, S., & Pasquale, B. J. (2017). Politicians, bureaucrats, and development: Evidence from india. *American Political Science Review*, 111(1), 162-183.
- Hahn, J., Todd, P., & Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1), 201-209.
- Hainmueller, J., Hall, A. B., & Snyder, J. M. (2015). Assessing the external validity of election rd estimates: An investigation of the incumbency advantage. *The Journal of Politics*, 77(3), 707-720.
- Hainmueller, J., Hangartner, D., & Pietrantuono, G. (2017). Catalyst or crown: Does naturalization promote the long-term social integration of immigrants? *American Political Science Review*, 111(2), 256-276.
- Hall, A. B. (2015). What happens when extremists win primaries? *American Political Science Review*, 109(1), 18-42.
- Hall, A. B., & Thompson, D. M. (2018). Who punishes extremist nominees? candidate ideology and turning out the base in us elections. *American Political Science Review*, 112(3), 509-524.
- Hidalgo, F. D., & Nichter, S. (2016). Voter buying: Shaping the electorate through clientelism. *American Journal of Political Science*, 60(2), 436-455.
- Hirano, S. (2011). Do individual representatives influence government transfers? evidence from japan. *The Journal of Politics*, 73(4), 1081-1094.
- Holbein, J. B. (2016). Left behind? citizen responsiveness to government performance information. *American Political Science Review*, 110(2), 353-368.
- Holbein, J. B., & Hillygus, D. S. (2016). Making young voters: The impact of preregistration on youth turnout. *American Journal of Political Science*, 60(2), 364-382.
- Hopkins, D. J. (2011). Translating into votes: The electoral impacts of spanish-language ballots. *American Journal of Political Science*, 55(4), 814-830.
- Imbens, G., & Kalyanaraman, K. (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3), 933-959.
- Kalšnja, M., & Titiunik, R. (2017). The incumbency curse: Weak parties, term limits, and unfulfilled accountability. *American Political Science Review*, 111(1), 129-148.



- Kelley, J. G., & Simmons, B. A. (2015). Politics by number: Indicators as social pressure in international relations. *American Journal of Political Science*, *59*(1), 55–70.
- Klašnja, M. (2015). Corruption and the incumbency disadvantage: Theory and evidence. *The Journal of Politics*, *77*(4), 928–942.
- Krasno, J. S., & Green, D. P. (2008). Do televised presidential ads increase voter turnout? evidence from a natural experiment. *The Journal of Politics*, *70*(1), 245–261.
- Larreguy, H., Marshall, J., & Querubín, P. (2016). Parties, brokers, and voter mobilization: How turnout buying depends upon the party’s capacity to monitor brokers. *American Political Science Review*, *110*(1), 160–179.
- Larreguy, H., Montiel Olea, C. E., & Querubin, P. (2017). Political brokers: Partisans or agents? evidence from the mexican teachers’ union. *American Journal of Political Science*, *61*(4), 877–891.
- Lerman, A. E., & McCabe, K. T. (2017). Personal experience and public opinion: A theory and test of conditional policy feedback. *The Journal of Politics*, *79*(2), 624–641.
- Lopes da Fonseca, M. (2017). Identifying the source of incumbency advantage through a constitutional reform. *American Journal of Political Science*, *61*(3), 657–670.
- Marshall, J. (2016). Education and voting conservative: Evidence from a major schooling reform in great britain. *The Journal of Politics*, *78*(2), 382–395.
- Mo, C. H., & Conn, K. M. (2018). When do the advantaged see the disadvantages of others? a quasi-experimental study of national service. *American Political Science Review*, *112*(4), 721–741.
- Mummolo, J. (2018). Modern police tactics, police-citizen interactions, and the prospects for reform. *The Journal of Politics*, *80*(1), 1–15.
- Nellis, G., & Siddiqui, N. (2018). Secular party rule and religious violence in pakistan. *American Political Science Review*, *112*(1), 49–67.
- Novaes, L. M. (2018). Disloyal brokers and weak parties. *American Journal of Political Science*, *62*(1), 84–98.
- Nyhan, B., Skovron, C., & Titiunik, R. (2017). Differential registration bias in voter file data: A sensitivity analysis approach. *American Journal of Political Science*, *61*(3), 744–760.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.
- Palmer, M., & Schneer, B. (2016). Capitol gains: The returns to elected office from corporate board directorships. *The Journal of Politics*, *78*(1), 181–196.
- Rozenas, A., Schutte, S., & Zhukov, Y. (2017). The political legacy of violence: The long-term impact of stalin’s repression in ukraine. *The Journal of Politics*, *79*(4), 1147–1161.
- Samii, C. (2013). Perils or promise of ethnic integration? evidence from a hard case in burundi. *American Political Science Review*, *107*(3), 558–573.
- Sances, M. W. (2017). Attribution errors in federalist systems: When voters punish the president for local tax increases. *The Journal of Politics*, *79*(4), 1286–1301.
- Schickler, E., Pearson, K., & Feinstein, B. D. (2010). Congressional parties and civil rights politics from 1933 to 1972. *The Journal of Politics*, *72*(3), 672–689.
- Sekhon, J. S., & Titiunik, R. (2012). When natural experiments are neither natural nor experiments. *American Political Science Review*, *106*(1), 35–57.
- Szakonyi, D. (2018). Businesspeople in elected office: Identifying private benefits from firm-level returns. *American Political Science Review*, *112*(2), 322–338.

- Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317.
- What Works Clearinghouse. (2017). *Procedures handbook* (Tech. Rep.). What Works Clearinghouse. (Version 4.0.)
- Xu, Y., & Yao, Y. (2015). Informal institutions, collective action, and public investment in rural china. *American Political Science Review*, *109*(2), 371-391.

# Supplement

## Contents

---

<b>S1 Sample selection</b>	<b>25</b>
S1.1 RD articles included . . . . .	25
S1.2 Excluded RD articles . . . . .	27
<b>S2 Replication data collection</b>	<b>30</b>
<b>S3 Yearly number of published RD studies</b>	<b>31</b>
<b>S4 Extraction of original t-statistics</b>	<b>32</b>
<b>S5 Additional caliper test results</b>	<b>34</b>
<b>S6 Reproduction of originally reported estimates</b>	<b>35</b>
<b>S7 P-values</b>	<b>36</b>
<b>S8 All studies with replication data</b>	<b>38</b>
<b>S9 Re-analysis using rdhonest package</b>	<b>39</b>

---

# S1 Sample selection

## S1.1 RD articles included

We collected our sample of 45 articles in two ways. First, we conducted a targeted search on Google Scholar within the three journals of interest. We searched for any articles mentioning terms which directly refer to the RD design (e.g. “regression discontinuity”) along with common phrases associated with RD estimators (e.g. “bandwidth” and “cut point”). Our second approach entailed searching for these terms on the journals’ websites and reading all articles’ abstracts from the past ten years.

We included in our sample articles which used an RD as the primary empirical strategy in an applied setting as well as studies where an RD complements another design (e.g., Broockman & Ryan, 2016). We excluded studies primarily making a methodological contribution (e.g., Cattaneo et al., 2016) and articles whose research settings are analogous to RD contexts but do not fit the standard definition of an RD. For example, some do not use a continuous running variable (e.g., Dunning & Nilekani, 2013 and de Kadt, 2017).

We classify as “primary” RD point estimates those which are referenced in the articles’ abstracts. As such, individual studies include multiple RD point estimates that we include in our sample. The one main estimate of each article was based on which of the primary estimates was most highlighted by the authors, or which one was most central to the authors’ key conclusion.

Among our sample of articles, we extracted salient information including the type of estimator used, bandwidth selection procedure (if applicable), number of units, point estimates, standard errors, and when explicitly reported,  $p$ -values. We list all articles’ authors, the publication, and RD type in Table S1. These studies span a broad range of substantive topics. For instance, Fourinaies and Hall (2014) use a sharp RD to study the link between incumbency and U.S. Congressional campaign contributions, and Szakonyi (2018) investigates whether a firm director’s electoral victory affects the firm’s profitability in the future. Cavaille and Marshall (2019) study how an additional year of schooling affects individuals’ attitudes towards immigration, and Boas, Hidalgo, and Richardson (2014) investigate how certain electoral outcomes affect government contracts.

**Table S1:** Sample of Political Science RD Studies/Outcomes

Author(s) & Year	Journal	RD type
Ariga (2015)	<i>JoP</i>	Sharp
Boas et al. (2014)	<i>JoP</i>	Sharp
Boas and Hidalgo (2011)	<i>AJPS</i>	Sharp
Bohlken (2018)	<i>AJPS</i>	Sharp
Brollo and Nannicini (2012)	<i>APSR</i>	Sharp
Broockman and Ryan (2016)	<i>AJPS</i>	Sharp
Carson and Sievert (2017)	<i>JoP</i>	Sharp
Caughey, Warshaw, and Xu (2017)	<i>JoP</i>	Sharp
Cavaille and Marshall (2019)	<i>APSR</i>	Fuzzy
Clinton and Sances (2018)	<i>APSR</i>	Sharp
Coppock and Green (2016)	<i>AJPS</i>	Fuzzy
Dahlgaard (2018)	<i>APSR</i>	Sharp
de Benedictis-Kessner (2018)	<i>JoP</i>	Sharp
de Benedictis-Kessner and Warshaw (2016)	<i>JoP</i>	Sharp
Eggers and Hainmueller (2009)	<i>APSR</i>	Sharp
Eggers and Spirling (2017)	<i>JoP</i>	Sharp
Erikson, Folke, and Snyder (2015)	<i>JoP</i>	Sharp
Ferwerda and Miller (2014)	<i>APSR</i>	Sharp
Fiva and Smith (2018)	<i>APSR</i>	Sharp
Folke, Persson, and Rickne (2016)	<i>APSR</i>	Sharp
Folke and Snyder (2012)	<i>AJPS</i>	Sharp
Fourinaies and Hall (2014)	<i>JoP</i>	Sharp
Galasso and Nannicini (2011)	<i>APSR</i>	Sharp
A. S. Gerber, Kessler, and Meredith (2011)	<i>JoP</i>	Sharp
E. R. Gerber and Hopkins (2011)	<i>AJPS</i>	Sharp
Gulzar and Pasquale (2017)	<i>APSR</i>	Sharp
Hainmueller, Hangartner, and Pietrantuono (2017)	<i>APSR</i>	Fuzzy
Hall (2015)	<i>APSR</i>	Sharp
Hall and Thompson (2018)	<i>APSR</i>	Sharp
Hidalgo and Nichter (2016)	<i>AJPS</i>	Sharp
Hirano (2011)	<i>JoP</i>	Sharp
Holbein (2016)	<i>APSR</i>	Fuzzy
Holbein and Hillygus (2016)	<i>AJPS</i>	Fuzzy
Klašnja (2015)	<i>JoP</i>	Sharp
Kalšnja and Tituiunik (2017)	<i>APSR</i>	Sharp
Larreguy, Marshall, and Querubín (2016)	<i>APSR</i>	Sharp
Lopes da Fonseca (2017)	<i>AJPS</i>	Sharp
Mo and Conn (2018)	<i>APSR</i>	Fuzzy
Novaes (2018)	<i>AJPS</i>	Sharp
Palmer and Schneer (2016)	<i>JoP</i>	Fuzzy
Rozenas, Schutte, and Zhukov (2017)	<i>JoP</i>	Fuzzy
Sances (2017)	<i>JoP</i>	Sharp
Schickler, Pearson, and Feinstein (2010)	<i>JoP</i>	Sharp
Szakonyi (2018)	<i>APSR</i>	Sharp
Xu and Yao (2015)	<i>APSR</i>	Sharp

## S1.2 Excluded RD articles

Some articles tentatively related to the RD design were excluded from our sample. This was made for primarily two reasons. First, studies were excluded if their research designs did not meet our definition of an RD design. For example, we do not consider studies with a clearly discrete running variable as an RD design for the purposes of this paper. Second, studies were excluded if their primary focus was to develop or refine existing RD estimators, and only used data from previous RD studies as illustration. We list all these excluded articles below.

Anzia and Berry (2011):

The RD-type analysis in their paper is not included as a primary component of their empirical analysis, and while they note that this approach is similar to an RD they do not classify it as one.

Cattaneo et al. (2016):

This is a methods-focused paper, deriving an estimator for a certain type of RD design.

Caughey, Dafoe, and Seawright (2017):

This is a methods-focused paper rather than an applied study.

Croke, Grossman, Larreguy, and Marshall (2016):

The authors note that their method is “similar to” an RD design but not exactly the same because their running variable is not continuous.

de Kadt (2017):

Study uses a discrete running variable, which is not consistent with our focus on implementing RD with continuous running variables.

Dunning and Nilekani (2013):

Despite being inspired by an RD design, it does not quite fit the criteria whereby units are assigned a value on a continuous running variable.

Eggers, Fowler, Hainmueller, Hall, and Snyder Jr. (2015):

This study mainly focuses on placebo tests and is a broader methodological discussion rather than an application.

Eggers (2017):

This study references different RD studies but does not implement an RD design on its own.

Eggers, Freier, Grembi, and Nannicini (2018):

This is a methods-focused paper exploring population-based threshold RD analyses. It is not an applied RD study.

Folke, Hirano, and Snyder (2011):

They employ an additional specification which focuses on close elections and they say is “similar in spirit” to RD designs. However, the authors explicitly note how their setting does not allow them to have the same causally identified estimate as an RD and thus do not consider their estimates to be RD estimates.

Friedman and Holden (2009):

The analysis does not incorporate the kind of continuous running variable which is standard for RD designs.

A. S. Gerber and Huber (2010):

This is too far from a standard RD design as it does not employ the typical assumption about continuity at the cut point.

Hainmueller, Hall, and Snyder (2015):

The authors are primarily interested in extrapolating RD treatment effects away from the cut point; not a typical RD design.

Hopkins (2011):

Excluded because there were such irregular and major spikes in the density of the running variable that it is behaving more like a discrete variable than a continuous one.

Kelley and Simmons (2015):

Their setting does not entail the running variable/cut point set-up necessary for a study to qualify as an RD design.

Krasno and Green (2008):

This paper refers to regression discontinuity designs for general motivation for their analysis, which is, in fact, a difference-in-differences estimator.

Larreguy, Montiel Olea, and Querubin (2017):

The study is primarily interested in an interaction between the RD treatment indicator and a background covariate, which puts the estimator/estimand outside the scope of our focus.

Lerman and McCabe (2017):

Study uses a discrete running variable, which is not consistent with our focus on implementing RD with continuous running variables.

Marshall (2016):

Study uses a discrete running variable, which is not consistent with our focus on implementing RD with continuous running variables.

Mummolo (2018):

Study uses time as a running variable, and so it's not consistent with our focus on running variables along which units receive a single score to the left or right of a discrete cut point.

Nellis and Siddiqui (2018)

This is an RD design combined with an instrumental variables designed in a way that adds several estimation and identification quirks. As such it doesn't neatly fit the definition of an RD study.

Nyhan, Skovron, and Titunik (2017):

Study uses a discrete running variable, which is not consistent with our focus on implementing RD with continuous running variables.

Samii (2013):

The study is primarily interested in an interaction between the RD treatment indicator and a background covariate, which puts the estimator/estimand outside the scope of our focus.

Sekhon and Titunik (2012):

This is a reanalysis of previous results and is methods-centric, rather than an applied paper.



## S2 Replication data collection

After generating the list of papers included in our study, we attempted to collect replication data for each one. We used publicly-available data whenever possible (e.g., replication materials made accessible through the Harvard Dataverse repository). Otherwise, we contacted the authors for replication data. We here list all studies for which we could not obtain replication data, along with an explanation of why the data are unavailable:

Brollo and Nannicini (2012):

The data were not available through an online repository or authors' websites. We contacted the authors multiple times but did not receive a reply.

Clinton and Sances (2018):

Some data necessary for the RD analyses are proprietary and were not available along with the replication materials.

Fiva and Smith (2018):

Data required for the RD analyses of interest are government sources and unavailable for public use. The data are not available online nor from the authors.

Folke et al. (2016):

Data required for the RD analyses of interest are government sources and unavailable for public use. The data are not available online nor from the authors.

Galasso and Nannicini (2011):

The data were not available through an online repository or authors' websites. We contacted the authors multiple times but did not receive a reply.

Hainmueller et al. (2017):

Data required for the RD analyses of interest are government sources and unavailable for public use. The data are not available online nor from the authors.

Hirano (2011):

Some data necessary for the RD analyses are proprietary and were not available along with the replication materials.

Holbein (2016):

Data required for the RD analyses of interest are government sources unavailable for public use. The data are not available online nor from the authors.

Mo and Conn (2018):

Data required for the RD analyses of interest are confidential and unavailable for public use. The data are not available online nor from the authors.

### S3 Yearly number of published RD studies

**Table S2:** Yearly Published RD Articles

Journal	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
APSR	1	0	1	1	0	1	2	3	3	7	<b>19</b>
AJPS	0	0	2	1	0	0	0	4	1	2	<b>10</b>
JOP	1	0	2	0	0	2	3	2	5	1	<b>16</b>
<b>All</b>	<b>2</b>	<b>0</b>	<b>5</b>	<b>2</b>	<b>0</b>	<b>3</b>	<b>5</b>	<b>9</b>	<b>9</b>	<b>10</b>	<b>45</b>

**Table S3:** Bandwidth Selection in Sample of RD Articles

Method	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
Automated (e.g. CCT and IK)	1	0	2	1	0	0	3	3	5	7	<b>22</b>
Non-automated	1	0	2	1	0	3	2	5	4	2	<b>20</b>
Global Polynomial	0	0	1	0	0	0	1	1	0	1	<b>4</b>
<b>All</b>	<b>2</b>	<b>0</b>	<b>5</b>	<b>2</b>	<b>0</b>	<b>3</b>	<b>6</b>	<b>9</b>	<b>9</b>	<b>10</b>	<b>46</b>

Note that Erikson et al. (2015) report as primary estimates of interest results derived from different bandwidth selection procedures. Therefore, we count this article twice in Table A3, once for each bandwidth selection, and so Table A3 includes one more unit (46) than Table A2 (45).

## S4 Extraction of original t-statistics

We collected point estimates, estimated standard errors, and  $p$ -values whenever they were explicitly reported in the paper. If a study did not report a  $p$ -value but reports the point estimate and standard error, we derived the implied  $p$ -value using a normal approximation of the sampling distribution. Some studies did not explicitly report standard errors or  $p$ -values. We imputed these values as described below. We refer to each article/estimate by the estimate code used in the dataset.

Ariga (2015) (a-b):

The standard error estimate was imputed by (a) subtracting the point estimate value from the upper value of the reported 90% confidence interval, and then (b) dividing that value by the critical value for a 90% confidence interval (1.64). A two-tailed, normal approximation  $p$ -value estimate was imputed using point estimate and imputed std. error because  $p$ -value was not explicitly listed in paper.

Caughey, Warshaw, and Xu (2017):

The standard error and  $p$ -value were imputed by running their replication code (which used `rdrobust`) and extracting the robust standard error value and robust  $p$ -value.

Carson and Sievert (2017):

The point estimate was rounded to the nearest tenth in text of original paper; imputed the full estimate that's derived by running the authors' replication code. The standard error and  $p$ -value were imputed by running their replication code and extracting the values.

Clinton and Sances (2018) (a-d):

We imputed the standard error using the width of the confidence interval. We imputed the  $p$ -value using the reported point estimate and the imputed standard error.

de Benedictis-Kessner (2018) (a-b):

The standard error was not explicitly listed, and so we imputed it by using the width of the reported confidence interval. The  $p$ -value explicitly listed in the main body of the article.

de Benedictis-Kessner and Warshaw (2016):

The standard error was not explicitly listed and so we imputed it by using the replication files made available for the paper. The  $p$ -value explicitly listed in the main body of the article.

Eggers and Spirling (2017) (a):

The point estimate was explicitly listed in main body of article, but note that there was a transcription error (the replication yields 1.911 instead of 0.911; the std. error value

was transcribed correctly). The standard error explicitly listed in main body of article, and a two-tailed, normal approximation  $p$ -value estimate was imputed by using the point estimate and std. error.

Ferwerda and Miller (2014):

The standard error was imputed by dividing the reported point estimate by the reported  $t$ -statistic. A two-tailed, normal approximation  $p$ -value estimate was imputed by using the reported point estimate and the imputed std. error.

Holbein (2016) (a-c):

The standard error was not explicitly listed and so we imputed by using the upper bound of a 95% CI and the point estimate along with the critical value of 1.96. A two-tailed  $p$ -value was not explicitly listed; we imputed the two-tailed  $p$ -value using the imputed SE value along with the point estimate.

Kalšnja and Titunik (2017):

The standard error was not explicitly listed and so we imputed by using the upper bound of a 95% CI and the point estimate along with the critical value of 1.96. However, the  $p$ -value was explicitly listed in main body of article.

Schickler et al. (2010) (a-d):

We retrieved the point estimate, standard error estimate, and  $p$ -value estimate by running the Stata replication files shared by authors.

## S5 Additional caliper test results

We ran the caliper test separately for studies that implemented an automated bandwidth selection procedure and those that used a non-automated approach to bandwidth selection. The results in Table S4a presents the imbalance around 1.96 for studies which used automated approaches to bandwidth selection, whereas Table S4b presents the results for studies which used non-automated approaches.

**Table S4:** Caliper tests by bandwidth selection method

(a) Automated BW selection method				(b) Other BW selection method			
Caliper	Below	Above	<i>p</i> -value	Caliper	Below	Above	<i>p</i> -value
10%	1	6	0.0625	10%	2	3	0.50
15%	1	8	0.0195	15%	3	3	0.66
20%	1	10	0.0059	20%	4	4	0.64

We also ran caliper tests for the t-statistics from the re-analysis. Table S5a presents the results for when the CCT bandwidth selection method is used, corresponding to Figure 3b in the main paper. Table S5b presents the results for when the CCT bandwidth selection method is used together with the bias adjustment and robust SEs, corresponding to Figure 3c in the main paper. Table S5 excludes Dahlgaard (2018) because their main estimate is a precision weighted estimate that was not part of our re-analysis.

**Table S5:** Caliper tests for re-analyzed data

(a) CCT bandwidth				(b) Bias corr. and robust SEs			
Caliper	Below	Above	<i>p</i> -value	Caliper	Below	Above	<i>p</i> -value
10%	2	4	0.34	10%	3	0	1
15%	3	4	0.50	15%	4	0	1
20%	3	4	0.50	20%	6	0	1

The caliper test is not a data-hungry test, so it will perform quite well with small samples. However, the samples are worrying small for these additional caliper tests, and power might be a concern. The results in this section should therefore be interpreted with caution. The tests were run (and reported) because it was requested during the review process.

## S6 Reproduction of originally reported estimates

This is section we document all failures to reproduce the authors' original estimates during our replication.

de Benedictis-Kessner and Warshaw (2016):

The authors used earlier version of `rdrobust`. When reproducing their estimates we select the bandwidth through the `rdbwselect_2014` for backward compatibility purposes, and the manually feed this bandwidth into the `rdrobust` package. The results are quite close but deviate slightly due to the slight differences in estimation code in `rdrobust` at the time the authors used it and now.

Eggers and Hainmueller (2009) (a) and (b):

There is a slight deviation between our reproduction results and those reported in the paper, though these differences are negligible. The replication data did not include code to allow us to precisely use the same functional form nor the same code to estimate robust standard errors.

Erikson et al. (2015) (a) through (c):

We received replication files from one author, and while we were able to get quite close with our attempts to reproduce the original estimates, there is a deviation between the ones reproduced and those originally reported in the paper. See STATA code “presidential elections reg and latex jms.do.”

E. R. Gerber and Hopkins (2011):

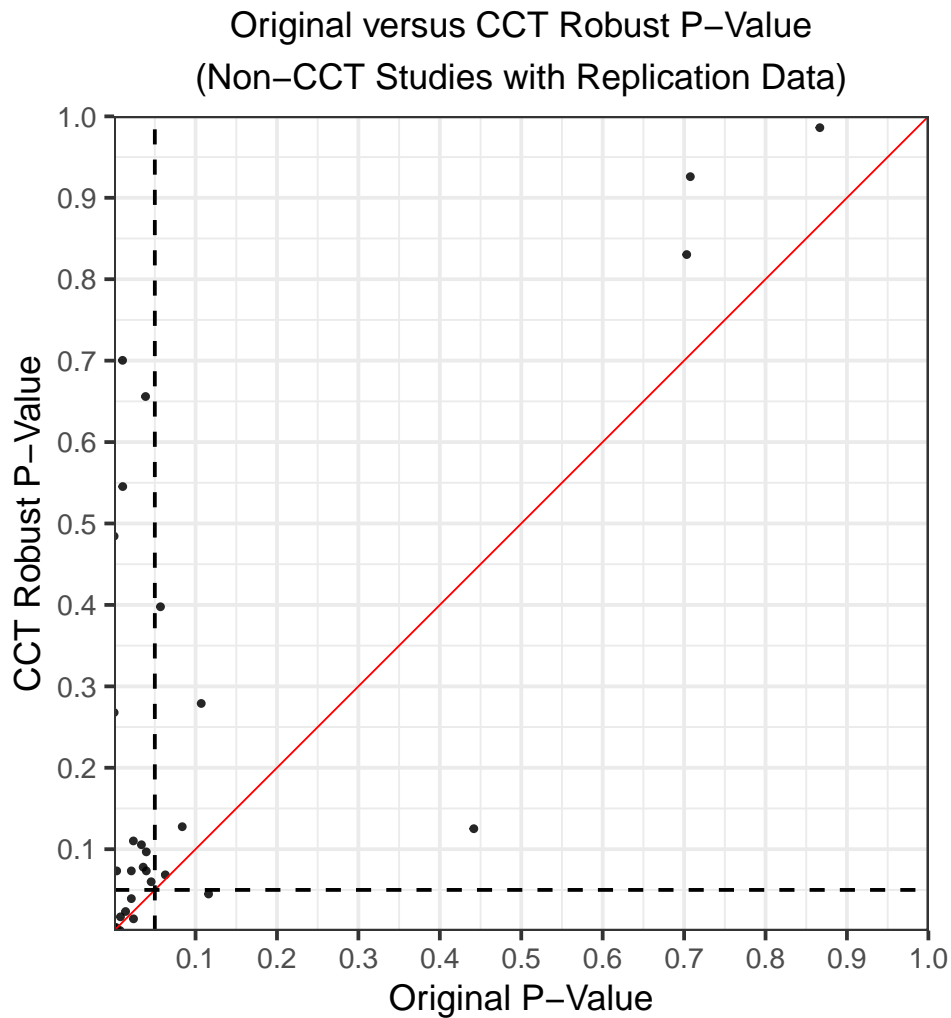
The authors noted that they used multiple imputation to fill in missing values for certain variables, though it wasn't exactly clear which ones. Analysis was run on the data that was available. Reanalysis results deviate from those in the paper.

Xu and Yao (2015):

We successfully reproduced the loess plots using their STATA code, which did not include code for producing the reported RD estimates. Using the same data, we implemented a loess estimation approach in R which produced a point estimate consistent with that reported in the paper.

## S7 P-values

The following two figures compare  $p$ -values reported in the original studies and the  $p$ -values for our re-analysis. If the original study did not report  $p$ -values, we have derived implied  $p$ -values as described in Section S4.



**Figure S1:** Original vs. CCT  $p$ -values for non-CCT studies

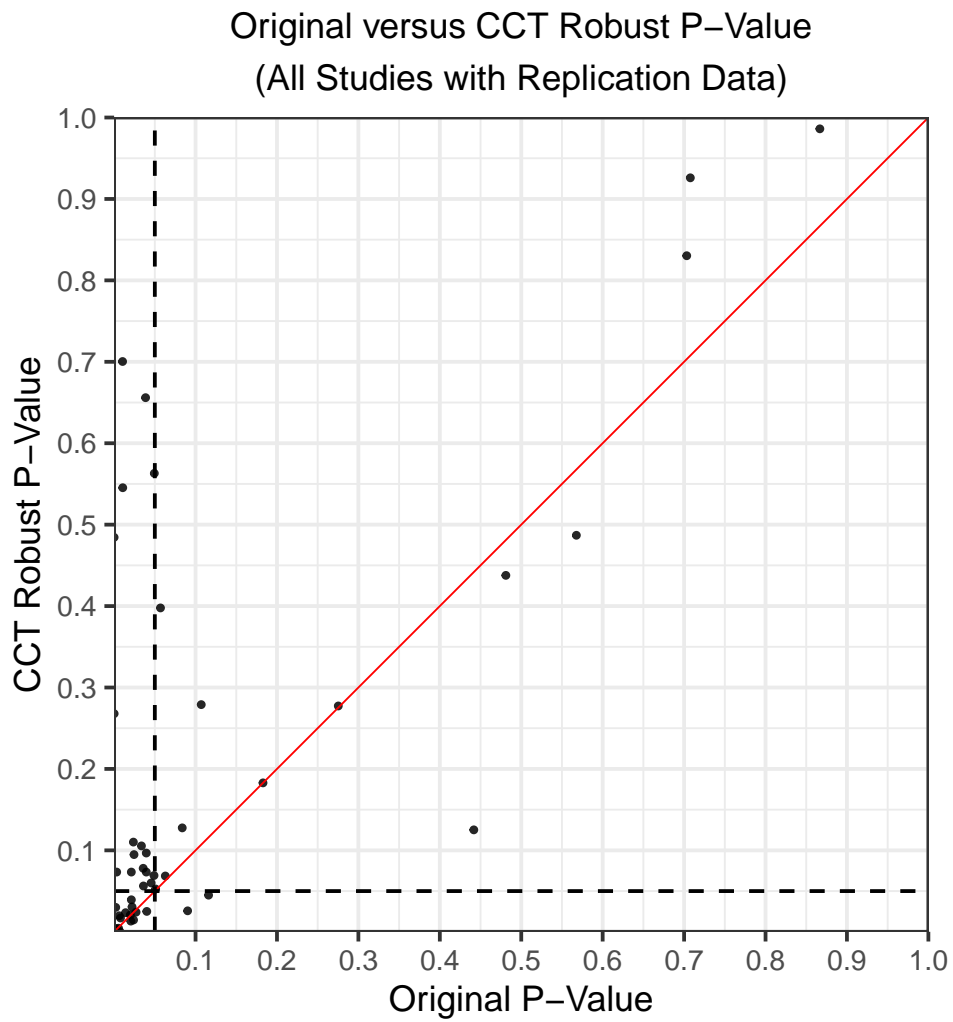
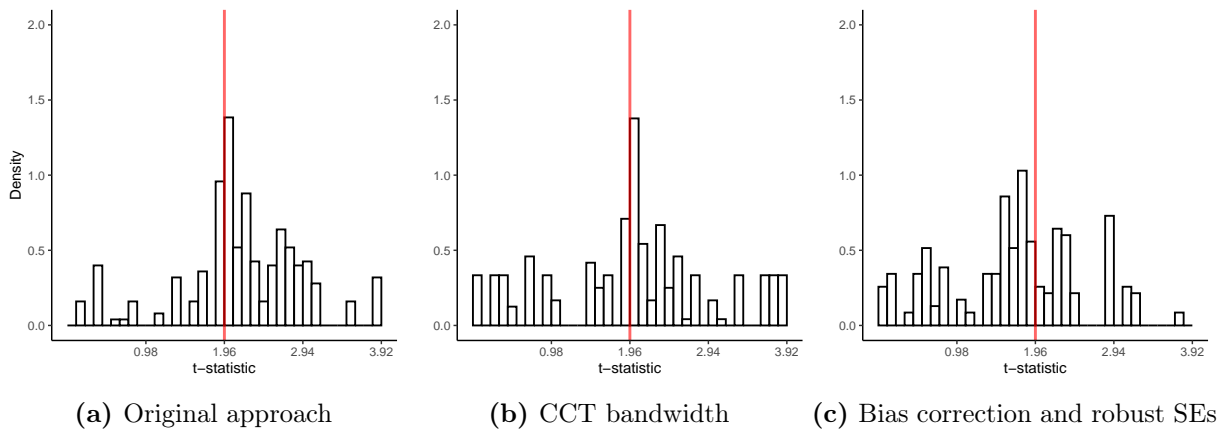


Figure S2: Original vs. CCT  $p$ -values for all studies

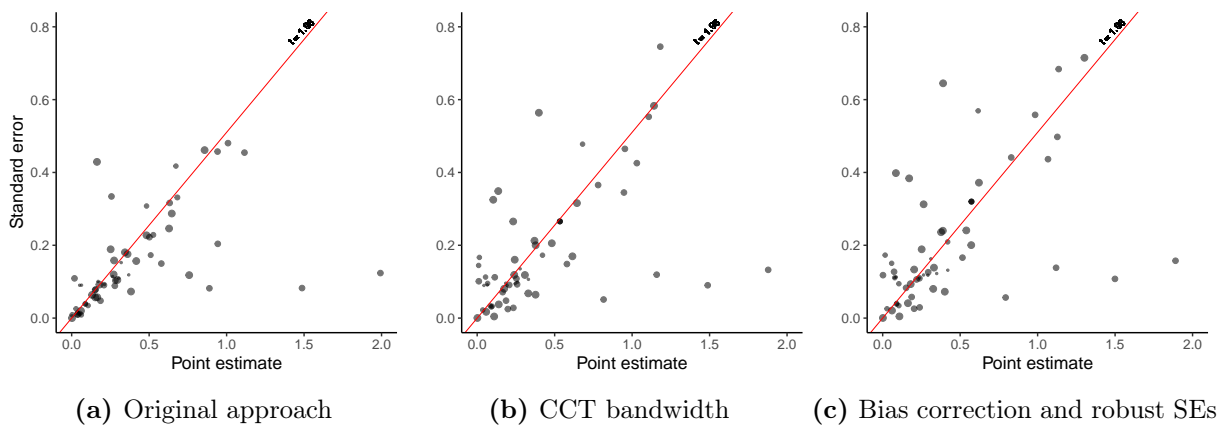


## S8 All studies with replication data

The following two figures correspond to Figures 3 and 4 in the main paper but includes all studies with replication data. That is, also studies that used the `rdrobust` package in their original analysis.



**Figure S3:** Distributions of  $t$ -statistics among replicated studies by method of analysis



**Figure S4:** Funnel plots for replicated studies by method of analysis

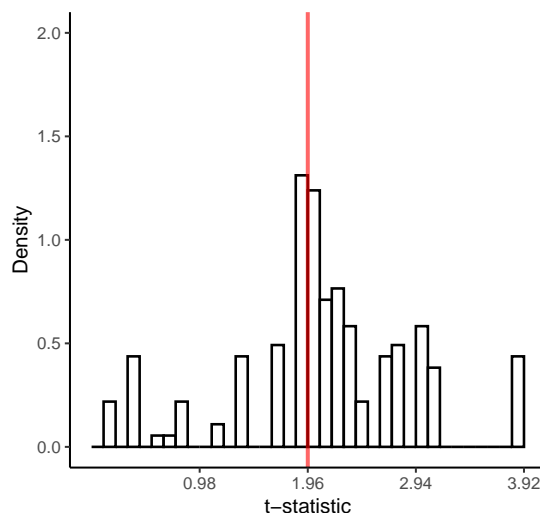
## S9 Re-analysis using rdhonest package

We also reanalyzed studies using the `rdhonest` package in R. To choose the bandwidth, we used a global quartic regression to estimate a bound on the second derivative for inference under second order Hölder class, as per the `NPR_MR0T.fit` function in `rdhonest`. The version of the `rdhonest` package we used cannot account for statistical clustering, and so we were limited to settings that did not originally use clustered standard errors. We construct pseudo  $t$ -statistics for the `rdhonest` package using the length of the confidence interval that include the bias adjustment.

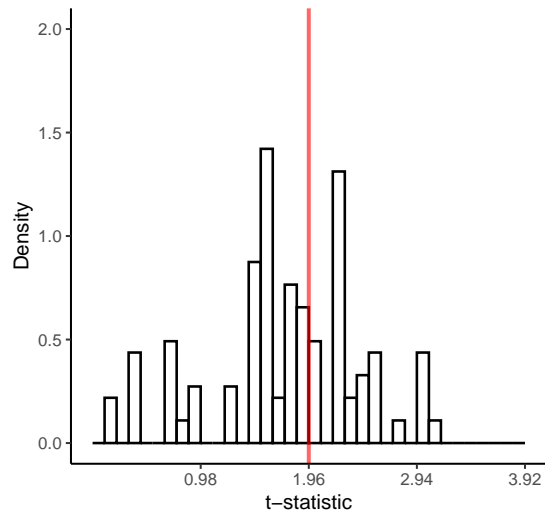
Figure S5 presents the original  $t$ -statistics for the studies we reanalyzed using `rdhonest`, and Figure S6 presents the distribution of  $t$ -statistics resulting from the `rdhonest` reanalysis. The reanalyses result in a leftward shift of these estimates.

Figures S7 and S8 are funnel plots similar to those in the main paper but using the `rdhonest` reanalysis. Figure S7 presents the original results for the subsample of studies included in the `rdhonest` reanalysis. Figure S8 presents the results from the `rdhonest` reanalysis.

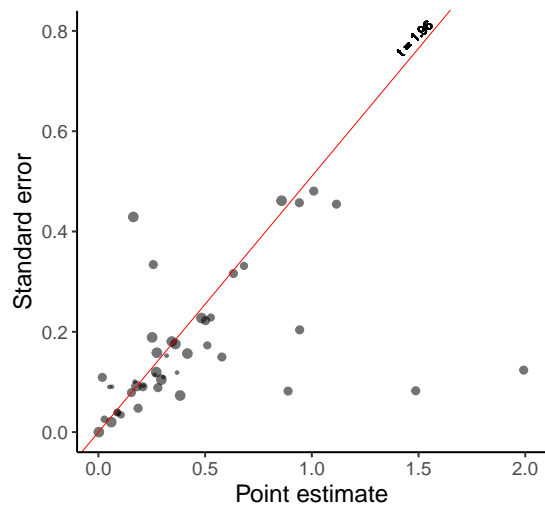
Finally, we compare the original, `rdrobust`, and `rdhonest`  $t$ -statistics in Table S6. This table is restricted to main estimates, which means that the study by Dahlggaard (2018) is excluded from rows 2 to 5 because their main estimate is a precision weighted estimate that was not part of our re-analysis. Both the `rdrobust` and `rdhonest` packages result in a left-ward shift of the  $t$ -statistics.



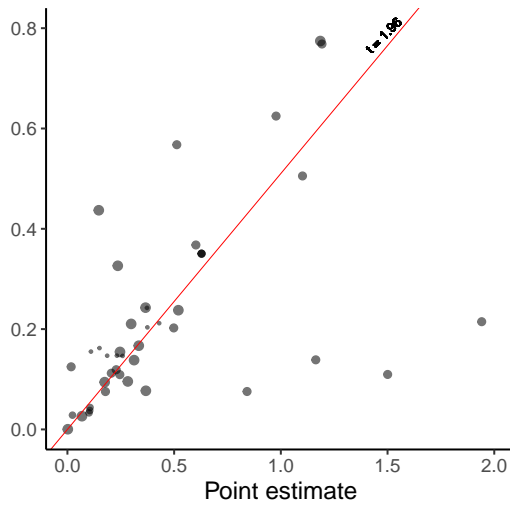
**Figure S5:** Original  $t$ -statistics for studies reanalyzed using RDHonest



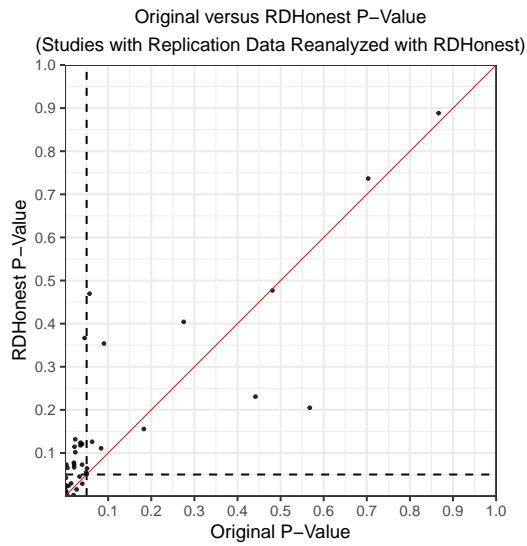
**Figure S6:** Reanalysis results: *t*-statistics for studies reanalyzed using RDHonest



**Figure S7:** Original funnel plot for studies reanalyzed using RDHonest



**Figure S8:** Reanalysis results: funnel plot for studies reanalyzed using RDHonest



**Figure S9:** Original  $p$ -values vs.  $p$ -values derived using RDHonest

**Table S6:** Original  $t$ -statistics versus `rdrobust` and `rdhonest`

Sample	Mean	S.D.	Quantiles						
			0.05	0.10	0.25	0.50	0.75	0.90	0.95
Original ests. (all papers)	4.11	7.03	0.39	1.44	2.00	2.26	3.52	5.00	14.18
Original (rdrobust studies)	4.59	7.90	1.05	1.66	1.98	2.26	3.33	5.96	16.62
Reanalysis results (rdrobust studies)	3.59	4.63	0.38	0.59	1.43	1.94	3.54	8.01	12.76
Original (RDHonest studies)	5.23	9.28	0.57	1.49	1.95	2.26	3.14	11.77	17.46
Reanalysis results (RDHonest studies)	3.00	3.25	0.41	0.80	1.53	1.85	2.58	6.94	8.91