Manipulation Tests in Regression Discontinuity Design: The Need for Equivalence Testing

Jack Fitzgerald, Vrije Universiteit Amsterdam*

July 31, 2024

Abstract

Researchers utilizing regression discontinuity design (RDD) commonly test for running variable (RV) manipulation around a cutoff, but incorrectly assert that insignificant manipulation test statistics are evidence of negligible manipulation. I introduce simple frequentist equivalence testing procedures that can provide statistically significant evidence that RV manipulation around a cutoff is practically equal to zero. I then demonstrate the necessity of these procedures, leveraging replication data from 36 RDD publications to conduct 45 equivalence-based RV manipulation tests. Over 44% of RV density discontinuities at the cutoff cannot be significantly bounded beneath a 50% upward jump. Bounding equivalence-based manipulation test failure rates beneath 5% requires arguing that a 350% upward density jump is practically equal to zero. Meta-analytic estimates reveal that average RV manipulation around the cutoff is equivalent to a 26% upward density jump. These results imply that many published RDD estimates may be confounded by discontinuities in potential outcomes due to RV manipulation that remains undetectable by existing tests. I provide research guidelines and commands in Stata and R to help researchers conduct more credible equivalence-based manipulation testing in future RDD research.

Keywords: McCrary density test, `rddensity`, `DCdensity`, Hartman test
JEL: C12, C18, C87, P00

*Email: j.f.fitzgerald@vu.nl. I thank Didier Nibbering for helpful comments and feedback. I currently hold a 12-month term as a member of the Superforecaster Panel for the Social Science Prediction Platform (SSPP; see DellaVigna, Pope, & Vivalt 2019). The views expressed in this paper do not necessarily represent the views of the SSPP, or of the researchers who created and/or operate the SSPP. This project has approval from the Ethics Review Board of Vrije Universiteit Amsterdam's School of Business and Economics.

1 Introduction

Regression discontinuity design (RDD) is one of the cornerstone quasi-experimental techniques that has propelled the credibility revolution in economics, political science, and other social sciences over the past 25 years (Imbens & Wooldridge 2009; Angrist & Pischke 2010; Samii 2016; Athey & Imbens 2017; Gopalan, Rosinger, & Ahn 2020, Imbens 2024). Cunningham (2021) finds that more than 5600 papers mentioning RDD were published in 2019 alone. RDD identifies local average treatment effects of interventions that are assigned when an agent's 'running variable' (RV) crosses some cutoff. Part of the reason for RDD's popularity is its 'experimental appeal'. For sufficiently granular RVs, people often trust that an agent's RV crossing the cutoff effectively randomizes that agent into or out of treatment, elminating selection biases that would arise if agents could choose their own treatment status.

This paper offers improvements on existing testing procedures that assess violations of a critical RDD identification assumption. Local average treatment effect identification in RDD hinges on the assumption that potential outcomes by treatment status are continuous functions of the RV as that RV crosses the cutoff. This assumption can be violated if agents with RV values near the cutoff endogenously manipulate their observed RV values to opt themselves into or out of treatment. If this kind of manipulation occurs, then RDD estimates will reflect not just treatment effects, but also confounding from differences in relevant characteristics between agents with different manipulation strategies (see Angrist & Pischke 2009; Lee & Lemieux 2010; Gerard, Rokkanen, & Rothe 2020; Cunningham 2021).

RV density discontinuity tests can assess *ex post* whether such RV manipulation around the cutoff has occurred. McCrary (2008) proposed the first such test, noting that if agents manipulate RV values around the cutoff, then this will be visible as a discontinuity in the RV's density at the cutoff. His `DCdensity` procedure estimates the one-sided limits of the RV's density as it approaches the cutoff from above and below, and assesses whether the (logarithmic) difference in RV density estimates at the cutoff is statistically significantly different from zero. An alternative version of this test, known as `rddensity`, has also recently emerged (Cattaneo, Jansson, & Ma 2018; 2020).

Such RV manipulation tests are quite popular in RDD papers. At time of writing, Web of Science reports that McCrary (2008) has around 1700 citations, and that Cattaneo, Jansson, & Ma (2018) and Cattaneo, Jansson, & Ma (2020) already have around 400 citations between them. These tests are a standard recommendation in texts on RDD and causal inference more generally, and are thus functionally required in RDD papers by journal editors and referees in economics, political science, and other disciplines (see Imbens & Lemieux 2008; Imbens & Wooldridge 2009; Lee & Lemieux 2010; Eggers et al. 2015; Athey & Imbens 2017; Cunningham 2021; Hartman 2021; Cattaneo & Titiunik 2022; Huntington-Klein 2022; Sieweki & Santoni 2022; Imbens 2024).

However, in practice, RV manipulation tests are usually applied fallaciously. Researchers utilizing RDD nearly always wish to demonstrate that RV manipulation near the cutoff is negligible. Researchers typically evidence this assertion by showing that RV density discontinuities at the cutoff are not statistically significantly different from zero (Hartman 2021). However, it is widely-known that this is bad scientific practice, as an underpowered estimate may be meaningfully large even if it is not statistically significantly different from zero (Altman & Bland 1995; Imai, King, & Stuart 2008 Wasserstein & Lazar 2016). Standard testing procedures may thus fail to detect meaningful RV manipulation near the cutoff.

This paper introduces equivalence testing frameworks that are more appropriate for demonstrating that RV manipulation around a cutoff is practically equal to zero. Under these frameworks, the researcher begins by setting a threshold for the maximally acceptable right-to-left ratio between the one-sided limits of the RV's density as it approaches the cutoff from each side. This is effectively a way of specifying the largest RV density discontinuity that would be 'economically insignificant'. The procedures then assess whether there is statistically significant evidence that the right-to-left ratio between these density estimates is significantly bounded within this threshold. I discuss two procedures that can provide such evidence. First, I discuss a novel equivalence testing approach to RV manipulation testing: the 'logarithmic density discontinuity' (LDD) equivalence test. Thereafter, I discuss Hartman's (2021) workhorse equivalence testing framework, upon which the LDD equivalence test is based and upon which my testing framework provides several improvements.

The social sciences are increasingly recognizing the need for equivalence testing when researchers wish to demonstrate null results or negligible relationships. Equivalence testing has a long history in biomedical sciences, and applications are being increasingly advanced in econometrics, psychology, and political science (e.g., see Piaggio et al. 2012; Hartman & Hidalgo 2018; Lakens, Scheel, & Isager 2018; Dette & Schumann 2024). Recent empirical work shows that up to 63% of estimates defending null claims made in top economics journals – roughly 90% of which are not statistically significantly different from zero – cannot be significantly bounded beneath reasonable benchmark effect sizes using equivalence testing (Fitzgerald 2024). This paper shows similar results for RV manipulation tests.

I demonstrate the necessity of equivalence testing for RV manipulation tests by showing that many RVs employed in published RDD papers fail lenient equivalence-based RV manipulation tests. My analysis draws on the replication data gathered by Stommes, Aronow, & Sävje (2023a), examining 45 RV density discontinuities at the cutoff across 36 RDD publications. Over 44% of these RVs' density discontinuities at the cutoff cannot be significantly bounded beneath a 50% upward jump (or equivalently, a 33.3% downward jump). To bring this 'failure rate' for equivalence-based RV manipulation tests beneath 5%, one must be willing to argue that a 350% upward jump in RV density at the cutoff is practically equal to zero. In fact, I show meta-analytic evidence that the average RV exhibits a density discontinuity equivalent to a 26% upward jump at the cutoff. These results suggest that in many published RDD analyses, estimated treatment effects may be confounded by meaningful RV manipulation at the cutoff that remains undetected by existing testing frameworks.

Given the clear need for equivalence-based approaches to RV manipulation tests, I conclude by providing guidelines on how such testing can be done credibly. I particularly advocate for researchers to set acceptable right-to-left density ratio thresholds idiosyncratically for each study, surveying independent experts about the smallest right-to-left density ratios that they would consider to be practically equal to zero (see also Fitzgerald 2024). I then provide commands in Stata and R that can be used to conduct such testing, including the `lddtest` command in Stata and the `lddtest` command in the `eqtesting` R package; Section 3.2 provides download instructions from Github. These procedures can provide more credible evidence that RV manipulation near the cutoff does not threaten the validity of treatment effects estimated by RDD.

2 Running Variable Manipulation Testing

Currently, researchers reporting RDD results predominantly use one of two tests to assess the presence of RV manipulation near the cutoff, the first of which is the `DCdensity` procedure in Stata and R (McCrary 2008).[1] The first and (historically) most popular RV manipulation test, `DCdensity` begins by creating a fine-gridded histogram of running variable Z and smoothing the histogram using separate local linear regressions to the left and right of cutoff c, respectively producing probability density function estimates $\hat{f}_-(Z)$ and $\hat{f}_+(Z)$. The command then computes the one-sided limits of Z's density as it approaches c from the left and right, which I respectively denote as $\hat{f}_-(c)$ and $\hat{f}_+(c)$. The estimate of interest to the `DCdensity` procedure is the logarithmic density discontinuity $\hat{\theta} \equiv \ln\left(\hat{f}_+(c)\right) - \ln\left(\hat{f}_-(c)\right)$. McCrary (2008) provides an asymptotically consistent standard error estimate for $\hat{\theta}$:

$$\text{SE}\left(\hat{\theta}\right) = \sqrt{\frac{24}{5Nh}\left(\frac{1}{\hat{f}_+(c)} + \frac{1}{\hat{f}_-(c)}\right)}. \tag{1}$$

Here N is the sample size and h is the total width of the bandwidth.[2] The procedure concludes by testing whether $\hat{\theta}$ is statistically significantly different from zero, assessing the extremity of test statistic $\frac{\hat{\theta}}{\text{SE}(\hat{\theta})}$ on the standard normal distribution.

The second density discontinuity testing procedure commonly used for RV manipulation testing is performed by the `rddensity` command in Stata, R, and Python (Cattaneo, Jansson, & Ma 2018; 2020), which differs from `DCdensity` in at least two important respects.

[1] `DCdensity` support is provided in Stata by the `DCdensity.ado` file hosted by Justin McCrary at https://eml.berkeley.edu/∼jmccrary/DCdensity/, and in R *via* the `rdd` R package (Dimmery 2016).

[2] For instance, if $\hat{f}_-(c)$ and $\hat{f}_+(c)$ are computed using observations with RV values in the range $[c-0.1, c+0.1]$, then $h = 0.2$. `DCdensity` has an automatic bandwidth selection procedure; see McCrary (2008).

First, though `DCdensity` and `rddensity` both estimate density functions $\hat{f}_-(Z)$ and $\hat{f}_+(Z)$, `rddensity` does so using local polynomial estimation rather than local linear histogram smoothing. Second, the `rddensity` procedure considers *linear* estimates of the RV's density discontinuity at the cutoff, rather than *logarithmic* estimates. Specifically, the point estimate of interest to `rddensity` is $\hat{f}_+(c) - \hat{f}_-(c)$. By default, `rddensity` computes an asymptotic plug-in estimator of $\widehat{\text{Var}}\left(\hat{f}_+(c) - \hat{f}_-(c)\right)$, though a jackknife alternative is available (Cattaneo, Jansson, & Ma 2018). The procedure also produces separate variance estimators for both RV density estimates to the left and right of the cutoff, respectively $\widehat{\text{Var}}\left(\hat{f}_-(c)\right)$ and $\widehat{\text{Var}}\left(\hat{f}_+(c)\right)$. `rddensity` concludes by assessing whether this linear RV density discontinuity is statistically significantly different from zero, examining the extremity of test statistic $\frac{\hat{f}_+(c)-\hat{f}_-(c)}{\sqrt{\widehat{\text{Var}}\left(\hat{f}_+(c)-\hat{f}_-(c)\right)}}$ on the standard normal distribution.

Though there are other estimators and tests of RV density discontinuities (see Otsu, Xu, & Matsushita 2013; Frandsen 2017; Bugni & Canay 2021; Ma, Jales, & Yu 2021; Igarashi 2023), these estimators and tests are not in widespread use. This is in part because apart from Bugni & Canay (2021), who offer the `rdcont` Stata package, no other proposed estimators or tests are accompanied by public-facing statistical software commands. However, even `rdcont` does not produce effect size estimates for the density discontinuity at the cutoff, and thus is not useful for the equivalence tests discussed in this paper.[3]

In practice, researchers are seldom interested in using these tests to demonstrate that there *is* evidence of RV manipulation around the cutoff. Rather, researchers ordinarily use these tests to demonstrate that such manipulation does *not* occur, and thus that endoge-

[3] As discussed in Section 3, an easily-interpretable effect size is functionally required for equivalence testing in this setting. This is because conducting equivalence testing requires one to be able to define a range of discontinuity values that are practically equal to zero, and to be able to test whether the estimated discontinuity is bounded within this range.

nous RV manipulation does not pose a threat to the validity of their causal identification strategy. To that end, researchers in practice interpret statistically insignificant manipulation test statistics as evidence that RV manipulation near the cutoff is negligible (Hartman 2021). Such practice is demonstrated in empirical applications both in the original publications introducing these methods (McCrary 2008; Cattaneo, Jansson, & Ma 2018; Cattaneo, Jansson, & Ma 2020) and in texts that advocate for the usage of these tests (e.g., see Lee & Lemieux 2010; Eggers et al. 2015; Cunningham 2021; Huntington-Klein 2022).

The common way in which these RV manipulation tests are used is thus inappropriate. As Cattaneo, Jansson, & Ma (2018; 2020) note, the hypotheses that are functionally assessed by both `DCdensity` and `rddensity` can be written as

$$\begin{aligned} H_0 &: \lim_{Z \to c^-} f(Z) = \lim_{Z \to c^+} f(Z) \\ H_A &: \lim_{Z \to c^-} f(Z) \neq \lim_{Z \to c^+} f(Z). \end{aligned} \tag{2}$$

Researchers typically use statistically insignificant test statistics in RV manipulation tests as evidence in favor of Equation 2's H_0. However, this inference is a well-known misinterpretation of statistical significance (see Altman & Bland 1995; Wasserstein & Lazar 2016).

As noted in Fitzgerald (2024), if the researcher uses the testing framework in Equation 2 when interested in showing that $\lim_{Z \to c^-} f(Z) = \lim_{Z \to c^+} f(Z)$, then the researcher begins by assuming in the null hypothesis that what they want to show is true. Such a testing framework shifts the burden of proof off of the researcher, who will not conclude that there is meaningful RV manipulation near the cutoff unless the data forces the researcher to abandon H_0. It is therefore a logical fallacy to infer that a statistically insignificant RV manipulation test result is evidence of no meaningful manipulation. Formally, this inference is an 'appeal to ignorance', which is committed when one argues that a claim is supported simply because no one has yet produced evidence against the claim. Imai, King, & Stuart (2008) specifically term this inference the 'balance testing fallacy' in placebo test contexts.

Thus in the way that they are currently applied, RV manipulation tests suffer from major credibility challenges. For researchers interested in showing that there is no RV manipulation near the cutoff, imprecision is 'good', in the sense that less power and more imprecision make it easier to obtain the researcher's desired finding (Imai, King, & Stuart 2008). This creates two perverse incentives. On one hand, simulation evidence shows that randomly dropping observations from a dataset can increase the likelihood of finding statistically insignificant placebo effects, even as the placebo effect estimates themselves grow larger (Imai, King, & Stuart 2008). Researchers can thus get closer to obtaining statistically insignificant RV manipulation test results either by trimming their sample or by setting restrictive bandwidths. On the other hand, statistically significant evidence of RV manipulation may go unreported. In some cases, there may even be good justification for this latter practice, as in very large datasets, a negligibly small RV density discontinuity may be misclassified as 'significant' simply due to very high power. Both of these selective reporting issues manifest as 'reverse p-hacking', and there is strong evidence of such selective reporting in top economics journals (Dreber, Johanneson, & Yang 2024). The key danger of these credibility flaws is that in many RDD studies, there may be meaningful RV manipulation near the cutoff that standard tests are not well-powered enough to detect, or significant RV manipulation near the cutoff that simply remains unreported to readers.

3 Equivalence-Based Manipulation Tests

A more credible equivalence-based testing framework for RV manipulation testing can be constructed from hypotheses of the form

$$\begin{aligned} H_0 &: \lim_{Z \to c^-} f(Z) \not\approx \lim_{Z \to c^+} f(Z) \\ H_A &: \lim_{Z \to c^-} f(Z) \approx \lim_{Z \to c^+} f(Z). \end{aligned} \tag{3}$$

If one can define a range of values for which $\lim_{Z \to c^-} f(Z) \approx \lim_{Z \to c^+} f(Z)$, then this is a feasibly testable hypothesis framework, as one can assess whether the RV's density discontinuity at the cutoff is bounded within that range using interval testing procedures. This section details two such procedures.

Both procedures that I detail here require the researcher to specify the largest ratio of $\hat{f}_+(c)$ to $\hat{f}_-(c)$ that they would consider 'practically equal' to 1. I parameterize this maximal acceptable right-to-left density ratio as $\epsilon > 1$. This effectively requires the researcher to specify the maximal RV density discontinuity that they would consider to be 'economically insignificant'. This is a subjective judgment call that will differ depending on the RV being examined and the specific research setting. In Section 6, I provide guidelines for how to credibly set this threshold.

Density ratios are useful effect size measures of RV density discontinuities at the cutoff because they are generally interpretable and comparable. Though some RV manipulation testing frameworks (e.g., `rddensity`) center linear density discontinuities as the estimand of interest, linear density discontinuities require idiosyncratic information about the dataset for proper interpretation, and are thus not generally comparable across datasets (Hartman

2021). For example, if Z crossing c induces a histogram discontinuity of 12 observations, this is more notable in a dataset of 100 observations than it is in a dataset of 1,000,000 observations. The usual practice of converting observation counts to probability densities creates similar comparability issues. For instance, if Z crossing c induces a three percentage point jump in probability density, this is more notable in a dataset of 1,000,000 observations than it is in a dataset of 100 observations. In contrast, density ratios at the cutoff are always comparable across datasets, and are relatively easy to interpret. This is important for equivalence-based testing procedures, as it helps researchers define valid thresholds for practically negligible effect sizes.

3.1 A Novel Testing Framework

`DCdensity` is particularly useful for addressing an unwieldy feature of tests on RV density ratios. Specifically, ratios are linearly asymmetric effect sizes in percentage terms. For example, a right-to-left density ratio of $\frac{3}{2}$ is equivalent to a 50% upward jump, whereas a right-to-left density ratio of $\frac{2}{3}$ is equivalent to a 33.3% downward jump. Fortunately, the *logarithms* of these right-to-left ratios are linearly symmetric around zero. For instance, $\ln\left(\frac{2}{3}\right) = -\ln\left(\frac{3}{2}\right)$. Estimates from `DCdensity` can thus be used to construct equivalence tests with linearly symmetric bounds. As aforementioned in Section 2, the point estimate of interest to the McCrary procedure can be written as

$$\begin{aligned}\hat{\theta} &\equiv \ln\left(\hat{f}_{+}(c)\right) - \ln\left(\hat{f}_{-}(c)\right) \\ &= \ln\left(\frac{\hat{f}_{+}(c)}{\hat{f}_{-}(c)}\right),\end{aligned}$$

and thus `DCdensity` directly estimates the logarithmic right-to-left ratio between RV density estimates at the cutoff. Further, McCrary (2008) shows that $\hat{\theta}$ is an asymptotically normal and consistent estimator of the RV's logarithmic density discontinuity at the cutoff, with a standard error that can be consistently estimated using the formula in Equation 1. This provides useful inference guarantees.

My proposed testing framework assesses whether the $\hat{\theta}$ estimate produced by `DCdensity` is bounded within a linearly symmetric interval that converts maximal acceptable right-to-left density ratio ϵ and its inverse $\frac{1}{\epsilon}$ into logarithms $\big(\ln(\epsilon)$ and $-\ln(\epsilon)$ respectively$\big)$. I term this framework the 'logarithmic density discontinuity (LDD) equivalence test.'

Definition 3.1 (The Logarithmic Density Discontinuity Equivalence Test)**.** *The researcher wants to test the hypotheses in Equation 3 with Type I error rate $\alpha \in (0, 0.5)$. They thus set maximally acceptable right-to-left density ratio $\epsilon > 1$, formulating null hypothesis*

$$H_0 : \ln\left(\lim_{Z \to c^+} f(Z)\right) - \ln\left(\lim_{Z \to c^-} f(Z)\right) < -\ln(\epsilon)$$

or

$$\ln\left(\lim_{Z \to c^+} f(Z)\right) - \ln\left(\lim_{Z \to c^-} f(Z)\right) > \ln(\epsilon)$$

and alternative hypothesis

$$H_A : \ \ln\left(\lim_{Z \to c^+} f(Z)\right) - \ln\left(\lim_{Z \to c^-} f(Z)\right) \geq -\ln(\epsilon)$$

and

$$\ln\left(\lim_{Z \to c^+} f(Z)\right) - \ln\left(\lim_{Z \to c^-} f(Z)\right) \leq \ln(\epsilon).$$

The researcher then estimates $\hat{\theta}$ *and* $SE\left(\hat{\theta}\right)$ *using the* **`DCdensity`** *procedure. Thereafter, test statistics are computed as*

$$t_{LDD}^{-} = \frac{\hat{\theta} - \ln(\epsilon)}{SE\left(\hat{\theta}\right)} \qquad\qquad t_{LDD}^{+} = \frac{\hat{\theta} + \ln(\epsilon)}{SE\left(\hat{\theta}\right)},$$

and the researcher obtains the relevant test statistic

$$t_{LDD} = \underset{t \in \left\{t_{LDD}^{-}, t_{LDD}^{+}\right\}}{\arg\min} \left\{|t|\right\}.$$

Let $\Phi(\cdot)$ *be the cumulative density function of the standard normal distribution, and let* $z_{\alpha}^{*} = \Phi^{-1}(1-\alpha)$. *If* $t_{LDD} = t_{LDD}^{-}$, *then the researcher rejects* H_0 *and concludes that* $\lim_{Z \to c^{-}} f(Z)$ *is practically equal to* $\lim_{Z \to c^{+}} f(Z)$ *if and only if* $t_{LDD} \geq z_{\alpha}^{*}$. *If* $t_{LDD} = t_{LDD}^{+}$, *then the researcher rejects* H_0 *and concludes that* $\lim_{Z \to c^{-}} f(Z)$ *is practically equal to* $\lim_{Z \to c^{+}} f(Z)$ *if and only if* $t_{LDD} \leq -z_{\alpha}^{*}$.

This test is an extension of the 'two one-sided tests' framework, a workhorse framework in frequentist equivalence testing (Schuirmann 1987). This test holds size α because its decision rule is based on the smaller of its two one-sided test statistics, and thus the test is an intersection-union test of two one-sided tests that each hold size α (Berger & Hsu 1996). The LDD equivalence test in Definition 3.1 can also be inverted to produce 'equivalence confidence intervals' (ECIs).

Definition 3.2 (The Logarithmic Density Discontinuity Equivalence Confidence Interval)**.** *The researcher wishes to assess the hypotheses in Equation 3 using a test with Type I error rate* $\alpha \in (0, 0.5)$. *They thus set a maximally acceptable right-to-left density ratio* $\epsilon > 1$, *obtain*

*$\hat{\theta}$ and $SE\left(\hat{\theta}\right)$ from the **`DCdensity`** procedure, and formulate a real interval*

$$ECI_{1-\alpha} = \left[\hat{\theta} - z_{\alpha}^{*} SE\left(\hat{\theta}\right), \hat{\theta} + z_{\alpha}^{*} SE\left(\hat{\theta}\right)\right].$$

The researcher concludes that $\lim_{Z \to c^-} f(Z)$ is practically equal to $\lim_{Z \to c^+} f(Z)$ if and only if $ECI_{1-\alpha} \subset [-\ln(\epsilon), \ln(\epsilon)]$.

When exponentiated, the LDD ECI provides the smallest range of ratios within which one can significantly bound the ratio of $\hat{f}_+(c)$ to $\hat{f}_-(c)$. One can conclude that the ratio of $\hat{f}_+(c)$ to $\hat{f}_-(c)$ is statistically significantly bounded between $\frac{1}{\epsilon}$ and ϵ if and only if the LDD ECI is entirely contained within $[-\ln(\epsilon), \ln(\epsilon)]$.

3.2 Statistical Software Commands

I provide statistical software commands in Stata and R to implement my proposed LDD equivalence testing procedure from Section 3.1. For Stata, I provide the `lddtest` command, which can be accessed from https://github.com/jack-fitzgerald/lddtest. For R, I provide a similar `lddtest` command in the `eqtesting` R package. `eqtesting` can be accessed from https://github.com/jack-fitzgerald/eqtesting. Each of these commands is effectively a wrapper for `DCdensity`. The Stata version of `lddtest` relies on McCrary's (2008) original Stata code, while the R version of `lddtest` uses the `DCdensity` command from Dimmery's (2016) `rdd` package as a dependency. In addition to the running variable, both commands require users to specify the cutoff c and the maximal acceptable right-to-left ratio ϵ between density estimates on each side of the cutoff.

3.3 The Hartman Test and `rddensity`

The LDD equivalence test draws inspiration from Hartman (2021), who establishes the current workhorse framework for equivalence-based RV manipulation testing. Her approach is similar in spirit to mine, but her framework relies on the linear RV density estimates at the cutoff $\hat{f}_+(c)$ and $\hat{f}_-(c)$ produced by `rddensity`, rather than the logarithmic density discontinuity estimate $\hat{\theta}$ produced by `DCdensity`. I term her framework the 'Hartman test'.[4]

Definition 3.3 (The Hartman Test)**.** *The researcher wishes to assess the hypotheses in Equation 3 using a test with Type I error rate $\alpha \in (0, 0.5)$. They thus set a maximally acceptable right-to-left density ratio $\epsilon > 1$ and formulate null and alternative hypotheses*

$$
\begin{aligned}
H_0 &: \frac{\lim\limits_{Z \to c^+} f(Z)}{\lim\limits_{Z \to c^-} f(Z)} < \frac{1}{\epsilon} \;\; or \;\; \frac{\lim\limits_{Z \to c^+} f(Z)}{\lim\limits_{Z \to c^-} f(Z)} > \epsilon \\
H_A &: \frac{\lim\limits_{Z \to c^+} f(Z)}{\lim\limits_{Z \to c^-} f(Z)} \geq \frac{1}{\epsilon} \;\; and \;\; \frac{\lim\limits_{Z \to c^+} f(Z)}{\lim\limits_{Z \to c^-} f(Z)} \leq \epsilon.
\end{aligned}
$$

*The researcher then uses the **`rddensity`** procedure to obtain estimates $\hat{f}_-(c)$, $\hat{f}_+(c)$, $\widehat{Var}\left(\hat{f}_-(c)\right)$, and $\widehat{Var}\left(\hat{f}_+(c)\right)$. Test statistics are computed as*

$$
t_H^- = \frac{\hat{f}_+(c) - \frac{\hat{f}_-(c)}{\epsilon}}{\sqrt{\widehat{Var}\left(\hat{f}_+(c)\right) + \frac{1}{\epsilon^2}\widehat{Var}\left(\hat{f}_-(c)\right)}} \qquad t_H^+ = \frac{\hat{f}_+(c) - \epsilon\hat{f}_-(c)}{\sqrt{\widehat{Var}\left(\hat{f}_+(c)\right) + \epsilon^2\widehat{Var}\left(\hat{f}_-(c)\right)}}, \tag{4}
$$

[4]The Hartman test can be implemented in R using the `rdd.tost.ratio` command, provided at https://github.com/ekhartman/rdd_equivalence/blob/master/RDD_equivalence_functions.R. The command relies on inputs that are generated by the `rddensity` command in R (Cattaneo, Jansson, & Ma 2020). See Hartman (2021) for details.

and the relevant test statistic is

$$t_H = \underset{t \in \left\{t_H^-, t_H^+\right\}}{\arg\min} \left\{|t|\right\}.$$

If $t_H = t_H^-$*, then the researcher rejects* H_0 *and concludes that* $\lim_{Z \to c^-} f(Z)$ *is practically equal to* $\lim_{Z \to c^+} f(Z)$ *if and only if* $t_H \geq z_\alpha^*$*. If* $t_H = t_H^+$*, then the researcher rejects* H_0 *and concludes that* $\lim_{Z \to c^-} f(Z)$ *is practically equal to* $\lim_{Z \to c^+} f(Z)$ *if and only if* $t_H \leq -z_\alpha^*$*.*

Hartman's testing framework can (at times) also be inverted to allow estimation of the smallest ratio value ϵ^* that would permit a statistically significant bounding of $\frac{\hat{f}_+(c)}{\hat{f}_-(c)}$. When tractable, this inversion procedure permits estimation of what I term the 'Hartman equivalence confidence interval' (Hartman ECI).[5]

Definition 3.4 (The Hartman Equivalence Confidence Interval)**.** *The researcher wishes to find the smallest ratio* $\epsilon^* > 1$ *such that one can significantly bound* $\frac{\hat{f}_+(c)}{\hat{f}_-(c)}$ *within the range* $[\frac{1}{\epsilon^*}, \epsilon^*]$ *at a significance level of* α *using the Hartman test in Definition 3.3. If* $\frac{\hat{f}_+(c)}{\hat{f}_-(c)} < 1$*, then the researcher solves*

$$z_\alpha^* = \frac{\hat{f}_+(c) - \frac{\hat{f}_-(c)}{\epsilon^*}}{\sqrt{\widehat{Var}\left(\hat{f}_+(c)\right) + \frac{1}{(\epsilon^*)^2}\widehat{Var}\left(\hat{f}_-(c)\right)}} \tag{5}$$

for ϵ^* *and selects the smallest* $\epsilon^* > 1$ *from among the quadratic solutions. If* $\frac{\hat{f}_+(c)}{\hat{f}_-(c)} > 1$*, then*

[5] In what follows, for simplicity, I omit the virtually nonexistent case where $\frac{\hat{f}_+(c)}{\hat{f}_-(c)} = 1$ exactly.

the researcher solves

$$-z_{\alpha}^{*} = \frac{\hat{f}_{+}(c) - \epsilon^{*}\hat{f}_{-}(c)}{\sqrt{\widehat{Var}\left(\hat{f}_{+}(c)\right) + (\epsilon^{*})^{2}\,\widehat{Var}\left(\hat{f}_{-}(c)\right)}} \tag{6}$$

for ϵ^{*} *and selects the smallest* $\epsilon^{*} > 1$ *from among the quadratic solutions.*

When the Hartman ECI is tractable, one can conclude that a density discontinuity at the cutoff is practically equal to zero under the Hartman test if and only if $\left[\frac{1}{\epsilon^{*}}, \epsilon^{*}\right] \subset \left[\frac{1}{\epsilon}, \epsilon\right]$. Because the Hartman ECI is just an inversion of the Hartman test, this decision rule produces identical conclusions to the Hartman test (conditional on the Hartman ECI being tractable). When defined, Hartman ECIs are asymmetric on the linear scale and symmetric on the logarithmic scale (Hartman 2021). The denominators of the test statistics in Equations 4, 5, and 6 are variants of the asymptotic plug-in variance estimator proposed for `rddensity` by Cattaneo, Jansson, & Ma (2020).

The LDD equivalence test that I propose in Section 3.1 improves on the Hartman test in two ways. First, density ratios are only a valid effect size measure if $\hat{f}_{-}(c)$ and $\hat{f}_{+}(c)$ both exceed zero. This condition should always hold in principle, as $\hat{f}_{-}(c)$ and $\hat{f}_{+}(c)$ are both point estimates of probability density functions. However, in practice, `rddensity` can frequently produce negative estimates of $\hat{f}_{-}(c)$ and/or $\hat{f}_{+}(c)$. As I note in Section 4, `rddensity` yields negative estimates of $\hat{f}_{-}(c)$ or $\hat{f}_{+}(c)$ for two of the 45 RVs in my replication sample.

`rddensity`'s capacity to produce non-positive probability density estimates is an issue of functional form misspecification. As noted in Section 2, `rddensity` estimates probability density functions using linearly additive polynomial functions of the RV. The resulting misspecification issue is closely related to a well-known problem of 'linear probability models' that regress outcomes bounded between zero and one on continuous predictors: such linear probability models can produce predicted probabilities that are greater than one or less than zero (Horrace & Oaxaca 2006). Though the local linear regression methods employed by `DCdensity` are not immune to this problem, `rddensity` amplifies this issue in its bias correction procedure, which relies on bias estimates from higher-order polynomial specifications (see Cattaneo, Jansson, & Ma 2018; 2020). This can lead to outliers with Z values far away from c effectively receiving undue weight in the probability density estimation (see Gelman & Imbens 2018). These properties pose validity challenges to the Hartman test, which relies on `rddensity`. If either $\hat{f}_{-}(c)$ or $\hat{f}_{+}(c)$ are non-positive, then the point estimate of interest to the Hartman test may be a negative ratio, require division by zero, or arise from comparisons of two negative point estimates that should in principle never be negative.

The LDD equivalence test addresses this issue by using `DCdensity` rather than `rddensity`. As aforementioned in Section 2, `DCdensity`'s estimates arise from local linear histogram smoothing, so its $\hat{\theta}$ estimator is effectively initialized on quantities that are non-negative by construction. Further, `DCdensity` approaches bias correction using undersmoothing, avoiding skew and inference issues associated with correcting for bias using estimates produced by higher-order polynomial specifications (McCrary 2008). Section 4 shows empirically that `DCdensity` is less likely than `rddensity` to yield non-positive $\hat{f}_{-}(c)$ or $\hat{f}_{+}(c)$ estimates.

The second key issue with the Hartman test is that the 'critical ratio' ϵ^* is not always tractable to calculate, even when $\hat{f}_{-}(c)$ and $\hat{f}_{+}(c)$ are both positive. This issue can arise from one of two scenarios. First, solving Equation 5 or Equation 6 yields quadratic solutions for

ϵ^*, producing solution candidates of the form

$$\epsilon^* = \frac{-v \pm \sqrt{v^2 - 4uw}}{2u}.$$

Online Appendix A shows that the radicands of these candidates are negative whenever

$$4\left(z_\alpha^*\right)^4 \mathrm{Var}\left(\hat{f}_-(c)\right) \mathrm{Var}\left(\hat{f}_+(c)\right) > \left(\hat{f}_-(c)\right)^2 + 4\left(z_\alpha^*\right)^2 \hat{f}_+(c)\mathrm{Var}\left(\hat{f}_-(c)\right).$$

If this occurs, then no analytic solution for ϵ^* exists on the real plane, as solving Equation 5 or Equation 6 for ϵ^* then requires taking the square root of a negative number. Second, even for non-negative radicands, there may be no ϵ^* candidate that exceeds 1, which is a required property of ϵ^* (see Definition 3.4). At times, the absence of an analytically tractable $\epsilon^* > 1$ can imply that there is no $\epsilon > 1$ for which one can significantly bound $\frac{\hat{f}_+(c)}{\hat{f}_-(c)} \in \left[\frac{1}{\epsilon}, \epsilon\right]$.

The LDD equivalence test that I propose in Section 3.1 does not suffer from this property. Provided that `DCdensity` can produce valid estimates for $\hat{\theta}$ and SE $\left(\hat{\theta}\right)$, it is always possible to obtain an ECI from the LDD equivalence test, which implies that one can always find a critical ϵ^* for which $\hat{\theta}$ is significantly bounded within $[-\ln(\epsilon^*), \ln(\epsilon^*)]$.

Resultantly, though no valid Hartman ECIs can be defined for many RV density discontinuities, similar issues do not arise for ECIs arising from my LDD equivalence test. As I discuss in Section 4, when RV density discontinuities are computed using local linear regression under `rddensity`, four of the 45 RVs in my replication sample have non-tractable Hartman ECIs. When local quadratic regression is used instead, `rddensity` produces non-tractable Hartman ECIs for seven RVs. However, I show that these tractability issues are completely eliminated for all RVs in my replication sample after switching to the LDD equivalence test.

4 Data and Methods

My empirical analysis leverages the replication data from Stommes, Aronow, & Sävje (2023a), who assess the robustness of RDD findings in top political science journals. Stommes, Aronow, & Sävje (2023a) systematically collect all empirical RDD articles published in *American Journal of Political Science*, *American Political Science Review*, and *Journal of Politics* from 2009-2018.[6] They obtain replication data on 36 publications, and make this data available in a Harvard Dataverse repository (Stommes, Aronow, & Sävje 2023b). Some of the publications which Stommes, Aronow, & Sävje (2023a) replicate store data in multiple datasets. I proceed by examining RV density discontinuities at the cutoff for each distinct dataset. This yields 45 RV manipulation tests across 36 RDD publications.

Though my results arise from data on political science publications, my findings are also very relevant for empirical practice in economics. The vast majority of my sample is comprised of data from close election designs, which remain quite popular in economics (see Cunningham 2021). 73% of the RVs in my sample are electoral vote shares, and 75% of the articles in my sample identify causal effects exclusively through the electoral victories that arise when these vote shares cross a given cutoff. Other publications in the sample exploit additional RVs that are popular in economic research, including spatial discontinuities and age discontinuities. In fact, of the 81 publications documented in Lee & Lemieux's (2010) survey of RDD applications in economics, nearly 42% exploit discontinuities in vote shares,

[6] One article with available replication data examined by Stommes, Aronow, & Sävje (2023a) is in fact published in print in 2019, but was published online in 2018.

spatial distance, and/or age for identification. Therefore, if robustness checks on the RVs in this sample reveal serious issues, then this raises concerns about the RVs used in many RDD applications in economics.

I estimate each RV density discontinuity in three ways. The first set of results arises from `DCdensity`. The second set is obtained by estimating the RV density discontinuity using `rddensity`, employing local linear estimation to obtain the point estimate and local quadratic estimation to compute the bias correction. The third set of results is again obtained from `rddensity`, but now utilizing local quadratic estimation for the point estimate and local cubic estimation for the bias correction. I restrict `rddensity` estimations to local linear and local quadratic specifications to avoid skew and inference issues that can arise when higher-order polynomials are estimated in RDD settings (see Gelman & Imbens 2018). In all cases, I estimate RV density discontinuities at the cutoff using the default bandwidths computed by each command. Though `DCdensity` produces a direct $\hat{\theta}$ point estimate, I can directly compute a point estimate of $\hat{\theta}$ for the linear `rddensity` estimates as $\hat{\theta} \equiv \ln\left(\hat{f}_{+}(c)\right) - \ln\left(\hat{f}_{-}(c)\right)$, provided that `rddensity` produces non-negative $\hat{f}_{-}(c)$ and $\hat{f}_{+}(c)$. I drop any `rddensity` estimates for which either $\hat{f}_{-}(c) \leq 0$ or $\hat{f}_{+}(c) \leq 0$. All estimates are computed in R.

Table 1 displays summary statistics of the results from my replications. Summary statistics are only calculated for 44 of the 45 discontinuities in my sample for variables concerning `rddensity`. This is because both local linear and local quadratic estimates from `rddensity` yield non-positive estimates for either $\hat{f}_{-}(c)$ or $\hat{f}_{+}(c)$ on one RV in the sample.

Sample sizes for the RVs themselves appear to be reasonably large. The median sample size N is 1450, and all RVs possess at least 134 observations. The mean of sample sizes is right-skewed by very large sample sizes near the top of N's distribution. This implies that in

	Min	P10	P25	P50	P75	P90	Max	Mean	SD	N
N	134	257.6	706	1450	21773	114514.6	517255	40928.244	102586.024	45
$\hat{\theta}$, `DCdensity`	-1.141	-0.273	-0.065	0.028	0.237	0.616	1.466	0.078	0.439	45
$\left\lvert\hat{\theta}\right\rvert$, `DCdensity`	0.000	0.016	0.044	0.140	0.375	0.784	1.466	0.279	0.345	45
SE $\left(\hat{\theta}\right)$, `DCdensity`	0.005	0.017	0.040	0.105	0.221	0.369	0.551	0.152	0.142	45
Standard p-value, `DCdensity`	0.000	0.000	0.001	0.236	0.556	0.824	0.999	0.321	0.33	45
ϵ^*, `DCdensity`	1.036	1.095	1.179	1.426	2.015	3.651	8.586	2.002	1.511	45
$\hat{\theta}$, `rddensity`, local linear	-2.223	-0.445	-0.103	0.034	0.474	0.785	1.389	0.086	0.575	44
$\left\lvert\hat{\theta}\right\rvert$, `rddensity`, local linear	0.000	0.036	0.064	0.235	0.529	0.822	2.223	0.391	0.427	44
Standard p-value, `rddensity`, local linear	0.000	0.010	0.102	0.340	0.492	0.791	1.000	0.363	0.291	44
ϵ^*, `rddensity`, local linear	1.066	1.112	1.402	1.984	5.028	9.896	45.882	4.891	7.792	41
$\hat{\theta}$, `rddensity`, local quadratic	-2.670	-0.308	-0.106	-0.017	0.118	0.514	1.848	0.002	0.568	44
$\left\lvert\hat{\theta}\right\rvert$, `rddensity`, local quadratic	0.000	0.030	0.062	0.107	0.315	0.608	2.670	0.295	0.483	44
Standard p-value, `rddensity`, local quadratic	0.000	0.021	0.162	0.541	0.669	0.878	1.000	0.478	0.315	44
ϵ^*, `rddensity`, local quadratic	1.114	1.137	1.295	1.511	2.096	3.323	25.390	2.726	4.221	38

Note: The N column denotes the number of RV density discontinuities for which the variable is non-missing.

Table 1: Summary Statistics

principle, most of the RV manipulation tests I perform should not be severely underpowered.

Some of the RV density discontinuities in this sample would be deemed significant by standard testing frameworks. Standard p-values are computed directly by `DCdensity` and `rddensity` using tests of the form in Equation 2. `rddensity` produces standard p-values beneath 5% for 15.6% of RV density discontinuities using local linear estimation, and for 11.1% of RV density discontinuities using local quadratic estimation. `DCdensity` estimates standard p-values below 5% for 37.8% of RV density discontinuities.

The median sizes of RV logarithmic density discontinuities at the cutoff range from 0.107 to 0.235, which correspond to upward RV density jumps of 11.3-26.5%. The ϵ^* thresholds displayed in Table 1 also show that rather large ϵ thresholds must be adopted to significantly bound most RV density discontinuities at the cutoff. Based on the estimates for `DCdensity`, one must set $\epsilon = 1.426$ to significantly bound even half of the RV density discontinuities at the cutoff in my sample. This requires arguing that a 42.6% upward jump in RV density at the cutoff is practically equal to zero. Even larger ϵ thresholds are required for similar boundings based on the estimates from `rddensity`.

Examining the observation counts for ϵ^* across each of the three testing frameworks makes the Hartman ECI's tractability issues apparent. As discussed in Section 3.3, when `rddensity`'s point estimates are computed using local linear (local quadratic) estimation, four (seven) RVs do not produce tractable Hartman ECIs. These cases arise either due to tractability failures of the Hartman ECI or because `rddensity` yields non-positive estimates for either $\hat{f}_{-}(c)$ or $\hat{f}_{+}(c)$.

I conduct equivalence testing for each RV density discontinuity by assessing whether there is statistically significant evidence at a 5% significance level that $\hat{\theta} \in [-\ln(1.5), \ln(1.5)]$. This functionally assesses whether each RV density discontinuity can be significantly bounded beneath a 50% upward jump (or equivalently, a 33.3% downward jump). I conduct this testing using the LDD equivalence test outlined in Definition 3.1 for the estimates derived from `DCdensity`, and do so using the Hartman test in Definition 3.3 for the estimates obtained from `rddensity`. In the terminology of these tests, my threshold effectively sets $\epsilon = 1.5$.

I select $\epsilon = 1.5$ as my benchmark right-to-left density ratio for three reasons. First, there is evidence from the epidemiology literature that an odds ratio of 1.5 corresponds closely to a Cohen's d value of 0.2, which is a small effect size per Cohen (1988). Setting $\epsilon = 1.5$ thus effectively assesses whether RV density discontinuities can be bounded beneath sizes typically judged to be small in the social sciences. Second, this follows the practice of Hartman (2021), who uses this threshold in her re-analysis of the vote share RVs constructed in Eggers et al. (2015). Third and finally, most people would likely find a 50% upward jump to be a large discontinuity in practical settings. For example, voters would be likely and rightfully be concerned about election results where the number of politicians just above the winning vote cutoff is 50% higher than the number of politicians just below.

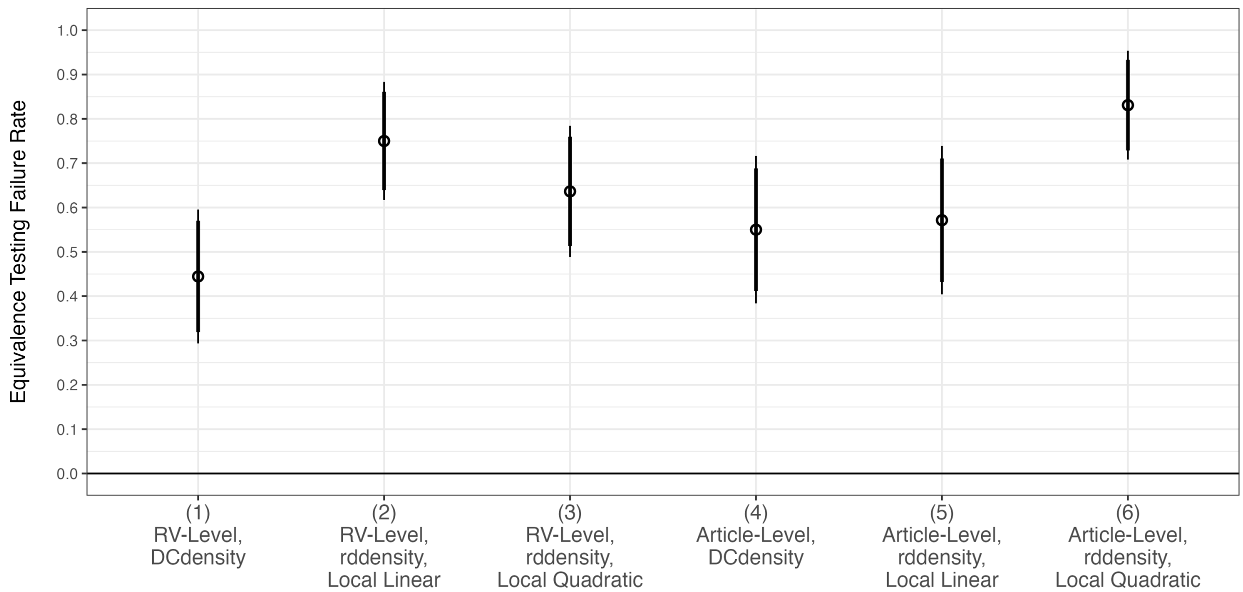
It should be easy to show that density discontinuities at the cutoff for RVs used for causal identification in RDD publications can be significantly bounded beneath a 50% upward jump, especially in high-caliber journals such as those sampled by Stommes, Aronow, & Sävje (2023a). If this condition holds for a given RV, then this RV 'passes' my lenient benchmark equivalence test. I compute the proportion of RVs that 'fail' this benchmark equivalence test, which I term the equivalence testing 'failure rate' (see also Fitzgerald 2024). I compute this failure rate at two levels. The RV-level failure rate is just the proportion of all RVs that fail the equivalence test. The article-level failure rate is obtained by first computing the RV-level failure rate *within* each article, and then calculating the average within-article failure rate *across* articles. I estimate precision by computing the standard errors of these means.

5 Results

5.1 Equivalence Testing Failure Rates

In my sample, equivalence testing failure rates for RV density discontinuities at the cutoff are exceptionally high. Figure 1 displays the main failure rate estimates, and Online Appendix Table A1 provides a table version of these results. These equivalence testing failure rates range from 44.4-83.1%. To provide a sense of interpretation, Model 1 in Figure 1 implies that in my sample, 44.4% of RV density discontinuities at the cutoff cannot be significantly bounded beneath a 50% upward jump. This implies that treatment effects estimated by RDDs that exploit the RVs in my sample may in many cases be confounded by meaningful endogenous RV manipulation near the cutoff, and there is no reliable evidence to reassure researchers that such manipulation does not occur.

These results cannot be explained by the choice of aggregation procedure; both RV-level and article-level failure rates are persistently significant. The results also cannot be explained away by the specific testing procedure employed. Equivalence testing failure rates remain significant regardless of whether logarithmic density discontinuities are estimated by `DCdensity` or `rddensity`, and regardless of whether local linear or local quadratic estimation is used to estimate the RV density discontinuity in `rddensity`.



Note: Equivalence testing failure rates at a 5% significance level with $\epsilon = 1.5$ are provided along with 90% and 95% confidence intervals based on the standard error of the mean. Equivalence tests are conducted using the LDD equivalence test in Definition 3.1 for results from `DCdensity` and using the Hartman test in Definition 3.3 for results from `rddensity`.

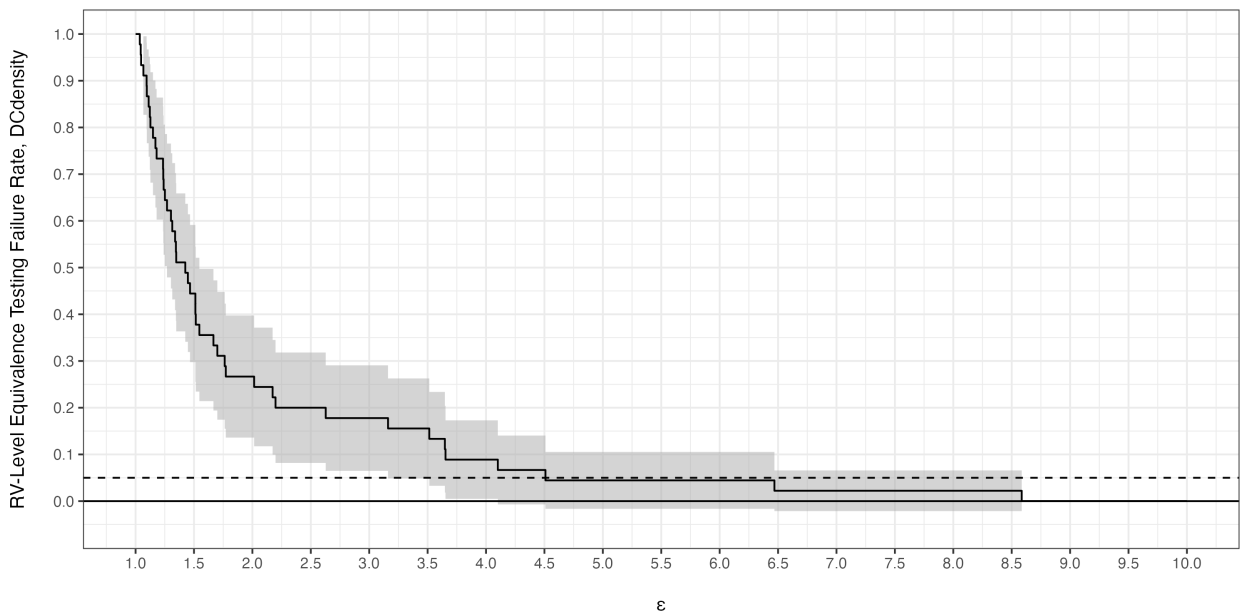
Figure 1: Main Failure Rate Estimates

Because these dimensions do not impact the significance of my main equivalence testing failure rate estimates, I primarily preference logarithmic RV density discontinuities arising from `DCdensity` in my discussion of results. I make this choice for two reasons. First, the RV-level failure rates arising from `DCdensity` are the smallest of my main failure rate estimates. Therefore, Model 1's estimates are the most robust in Figure 1 given my general findings about the severe magnitude of equivalence testing failure rates. Second, as discussed in Section 2, `DCdensity` produces a standard error for logarithmic density discontinuities, whereas this is not true for `rddensity`; the failure rates displayed in Figure 1 must be computed using the Hartman test (see Section 3.3). This point gains hightened importance when I compute meta-analytic estimates for logarithmic RV density discontinuities in Section 5.3.

5.2 Failure Curves

My large equivalence testing failure rate estimates cannot be explained by my choice of maximal acceptable right-to-left density ratio ϵ. Figure 2 displays the 'failure curve' for `DCdensity`, which shows the distribution of RV-level equivalence testing failure rates across different ϵ thresholds (see also Fitzgerald 2024). The shape of the failure curve reflects the fact that equivalence testing failure rates decline when one is willing to deem larger RV density discontinuities at the cutoff to be 'practically equal to zero'.

The failure curve in Figure 2 shows that in my sample, equivalence testing failure rates for RV manipulation tests remain significantly above nominal levels even as the ϵ threshold is allowed to grow exceedingly large. Consider the RV density discontinuity that one would need to tolerate in order to obtain equivalence testing failure rates beneath a traditional 5%



Note: The failure curve for `DCdensity` is displayed with uncertainty bands representing 95% confidence intervals of RV-level equivalence testing failure rates, based on the standard error of the mean. Equivalence testing failure rates are computed based on the LDD equivalence test in Definition 3.1. The solid and dashed horizontal lines respectively denote 0% and 5% failure rates.

Figure 2: Failure Curve

rate. The horizontal dashed line in Figure 2 denotes a 5% failure rate; the failure curve only crosses this line when $\epsilon = 4.509$. This implies that in order to obtain equivalence testing failure rates beneath 5%, one would need to be willing to claim that a right-to-left density ratio of $\epsilon = 4.509$ is practically equal to 1. This is identical to arguing that a 350.9% upward jump in RV density at the cutoff is practically equal to zero. Because such an argument is ludicrous, a more reasonable alternative conclusion emerges: RV manipulation near the cutoff cannot be reliably bounded beneath reasonable thresholds for a substantial proportion of RVs used in RDD analyses published in top political science journals.

Online Appendix Figure A1 shows these failure curves for each testing procedure. The headline results for the failure curve in Figure 2 hold for all testing procedures examined in Figure A1. In fact, failure rates arising from `DCdensity` are stochastically dominated by those from `rddensity` for all ϵ thresholds, confirming that `DCdensity` produces the lowest failure rates for equivalence-based RV manipulation tests in my sample.

5.3 Meta-Analytic Results

How large is the average RV density jump at the cutoff? To examine this question, I obtain a meta-analytic estimate of *absolute* logarithmic RV density discontinuities, utilizing LDD estimates $\hat{\theta}$ and their standard errors $\text{SE}\left(\hat{\theta}\right)$ from `DCdensity`. I focus on absolute LDDs rather than raw LDDs because RV manipulation in either direction of the cutoff raises concerns about treatment effects estimated by RDD.

I compute my meta-analytic estimate using an unrestricted weighted least squares approach (Ioannidis, Doucouliagos, & Stanley 2022; Stanley et al. 2023). Let i index a given

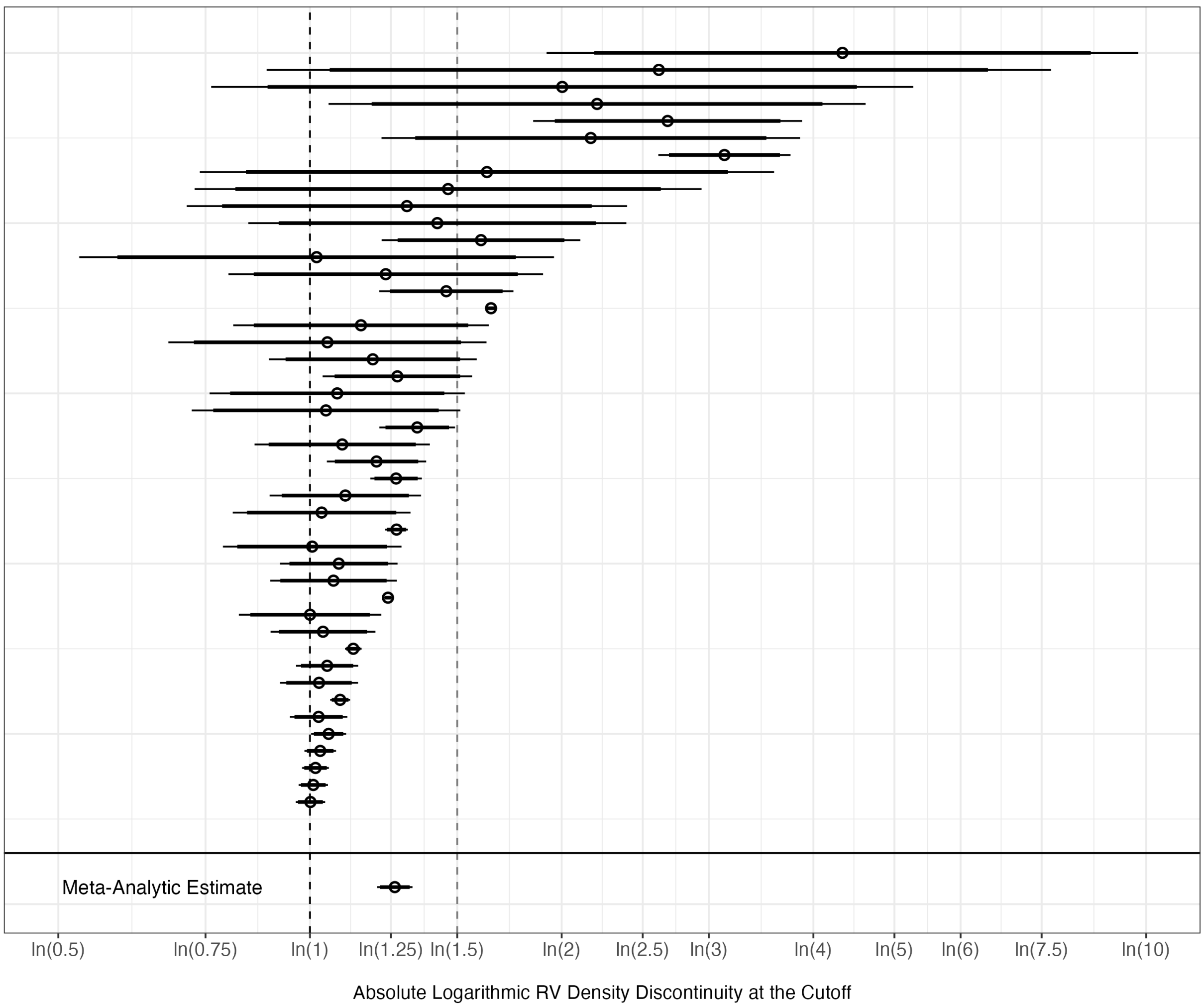
RV. This approach employs a regression of the form

$$\frac{\left|\hat{\theta}_i\right|}{\mathrm{SE}\left(\hat{\theta}_i\right)} = \beta \frac{1}{\mathrm{SE}\left(\hat{\theta}_i\right)} + \mu_i,$$

where β is the meta-analytic estimate of interest. In this setting, β is equivalent to a weighted average of effect sizes $\left|\hat{\theta}_i\right|$, where the weights are given by $\left(\mathrm{SE}\left(\hat{\theta}_i\right)\right)^{-2}$ (Stanley et al. 2023). This reflects the fact that unrestricted weighted least squares gives more weight to more precise estimates, which gives the procedure strong empirical advantages over other meta-analytic estimation methods such as random effects estimation (see Stanley et al. 2023).

Figure 3 displays the meta-analytic absolute LDD estimate, along with absolute LDDs at the cutoff for each RV examined in my sample. The meta-analytic estimate for $\left|\hat{\theta}\right| = 0.234$, with a standard error $\mathrm{SE}\left(\left|\hat{\theta}\right|\right) = 0.025$. This estimate implies that the meta-analytic average RV density discontinuity at the cutoff is equivalent to a 26.3% upward jump. This average discontinuity is precisely estimated, and is quite significantly bounded beneath a threshold of ln(1.5). However, as aforementioned in Section 4, I select this threshold in part because it is particularly large, and would raise valid manipulation concerns in research-relevant settings such as elections. For many RDD applications, researchers may have good reason to be wary of RV manipulation of the magnitude demonstrated in Figure 3.

These results contradict findings from prior research that conducts RV manipulation tests on large sets of RVs. In particular, Eggers et al. (2015) find that pooling 20 international samples of vote share RVs yields a logarithmic RV density discontinuity estimate that is not statistically significantly different from zero. Hartman's (2021) reproduction of Eggers et al. (2015) also shows that this pooled estimate is quite small, corresponding to just a 2% RV



Note: Above the black horizontal line, absolute logarithmic RV density discontinuities at the cutoff are displayed for each of the 45 RVs in my sample, along with 90% and 95% confidence intervals. The meta-analytic estimate is displayed below the black horizontal line. Dashed vertical black and gray lines respectively denote zero and ln(1.5). Logarithmic RV density discontinuities are estimated using `DCdensity`.

Figure 3: Meta-Analytic Estimates

density jump at the cutoff with a Hartman ECI of $[0.95, 1.05]$.

The much larger meta-analytic RV density discontinuity in my sample arises largely because my meta-analytic estimate and the pooled estimate from Eggers et al. (2015) are estimating two different parameters that have two different interpretations. Whereas my meta-analytic estimate shows the average *absolute* LDD in my sample, the logarithm of the pooled estimate from Hartman's (2021) reproduction of Eggers et al. (2015) provides the average *raw* LDD in the sample from Eggers et al. (2015). The pooled estimate in Eggers et al. (2015) thus does not rule out meaningful vote share manipulation, as this pooled estimate is consistent with some elections exhibiting vote share manipulation towards victory and some exhibiting vote share manipulation towards defeat, with the grand mean of such manipulation averaging out to zero. In contrast, my meta-analytic estimate shows that when looking at one RV at a time, RV manipulation at the cutoff tends to differ quite substantially from zero for each RV, regardless of whether it is RV manipulation into or out of treatment. Taking the average of these absolute LDDs shows that the average RV has some manipulation around the cutoff that is statistically significantly different from zero, regardless of direction.

6 Conclusion

I introduce several equivalence-based RV manipulation tests for RDD applications. In a large sample of RDD publications in top political science journals, I find that RVs often fail lenient versions of these tests. Over 44% of RVs in these publications have density discontinuities at the cutoff that cannot be significantly bounded beneath a 50% upward jump. Bringing equivalence testing failure rates beneath 5% requires arguing that upward

RV density jumps of 350% are practically equal to zero, and meta-analytic estimates even suggest an average upward RV density jump of 26% at the cutoff. These results suggest that many RVs used in RDD research may exhibit meaningful manipulation near the cutoff that standard testing procedures cannot detect. Therefore, in many RDD publications, treatment effect confounding from RV manipulation near the cutoff cannot be reliably ruled out.

The empirical findings in this paper mirror those in Hartman's (2021) re-analyses of RV manipulation tests for vote share RVs. Eggers et al. (2015) conduct RV manipulation tests on 20 electoral vote share RVs, finding that 95% produce estimates that are not statistically significantly different from zero. Hartman (2021) re-analyzes these RVs using equivalence-based RV manipulation tests (particularly using `rddensity` and the Hartman test), finding that 35% of RV density discontinuities at the cutoff cannot be significantly bounded beneath a 50% upward jump. The equivalence testing failure rates that I find in this paper are even larger than those found by Hartman (2021), remain high across a range of different testing specifications, and are demonstrated for a broader range of popular RV types that are frequently applied in published RDD papers in economics and political science.

Because these findings make clear that equivalence-based procedures are needed for RV manipulation testing in RDD, I conclude by offering guidelines on how such testing can be done credibly. Perhaps the most important question is the maximally acceptable right-to-left ratio ϵ between RV density limits at the cutoff. This threshold is ultimately a subjective judgment call, and will differ for different research settings. Thus though benchmark thresholds are useful in meta-analytic work that examines research across an entire field, the same cannot be said for individual RDD applications. The $\epsilon = 1.5$ threshold that I use in this paper is by no means a universal benchmark for all studies, and is selected for this analysis in part because it represents a discontinuity that would be seen as very large in practice (see Section 4). I therefore recommend setting ϵ idiosyncratically for each unique research setting, rather than relying on disciplinary benchmark thresholds.

Credible equivalence testing depends on the threshold ϵ being set independently to avoid p-hacking (Campbell & Gustafson 2021; Fitzgerald 2024). To that end, I recommend that researchers aggregate ϵ by surveying other experts for their judgments of the largest percentage upward jump in RV density at the cutoff that they would consider to be practically equal to zero. Online platforms such as the Social Science Prediction Platform (DellaVigna, Pope, & Vivalt 2019) possess centralized pools of researchers who can offer their judgments and predictions on the RV density discontinuity that will be observed at a given cutoff. If the research setting is highly idiosyncratic and requires specialized field knowledge for reasonably accurate predictions and judgments, then it may be reasonable to use the platform's email list feature to invite specific groups of researchers with specialized field expertise to offer their predictions and judgments. Though researchers may reasonably consider such a survey to be too much of a time and effort investment to elicit a threshold for a robustness check, eliciting predictions and judgments on RV density discontinuities can be naturally coupled with eliciting predictions and judgments on the primary treatment effect(s) of interest, which can yield a myriad of useful insights (see DellaVigna, Pope, & Vivalt 2019).

Once a maximally acceptable right-to-left density ratio ϵ is set, I recommend using my proposed logarithmic density discontinuity equivalence testing approach to assess whether the right-to-left ratio of density estimates on each side of the cutoff can be significantly bounded beneath ϵ (see Section 3.1). I recommend using this approach, rather than combining `rddensity` with the Hartman test, to avoid invalid non-positive density estimates that can emerge from `rddensity` and intractible test results that can arise from the Hartman test (see Section 3.3 and Section 4). My recommended procedure can be implemented using the `lddtest` Stata command or the `lddtest` command in the `eqtesting` R package. Both the `eqtesting` R package and the `lddtest` Stata command are available for download from Github, and Section 3.2 provides download instructions.

References

Altman, D. G. and J. M. Bland (Aug. 1995). "Statistics notes: Absence of evidence is not evidence of absence". *BMJ* 311.7003, pp. 485–485. DOI: 10.1136/bmj.311.7003.485.

Angrist, Joshua D and Jörn-Steffen Pischke (May 2010). "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics". *Journal of Economic Perspectives* 24.2, pp. 3–30. DOI: 10.1257/jep.24.2.3.

— (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Athey, Susan and Guido W. Imbens (May 2017). "The state of applied econometrics: Causality and policy evaluation". *Journal of Economic Perspectives* 31.2, pp. 3–32. DOI: 10.1257/jep.31.2.3.

Berger, Roger L. and Jason C. Hsu (Nov. 1996). "Bioequivalence trials, intersection-union tests and equivalence confidence sets". *Statistical Science* 11.4. DOI: 10.1214/ss/1032280304.

Bugni, Federico A. and Ivan A. Canay (Mar. 2021). "Testing continuity of a density via g-order statistics in the regression discontinuity design". *Journal of Econometrics* 221.1, pp. 138–159. DOI: 10.1016/j.jeconom.2020.02.004.

Campbell, Harlan and Paul Gustafson (Feb. 2021). "What to make of equivalence testing with a post-specified margin?" *Meta-Psychology* 5. DOI: 10.15626/mp.2020.2506.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma (Mar. 2018). "Manipulation testing based on density discontinuity". *The Stata Journal* 18.1, pp. 234–261. DOI: 10.1177/1536867x1801800115.

— (Sept. 2020). "Simple local polynomial density estimators". *Journal of the American Statistical Association* 115.531, pp. 1449–1455. DOI: 10.1080/01621459.2019.1635480.

Cattaneo, Matias D. and Rocío Titiunik (Aug. 2022). "Regression discontinuity designs". *Annual Review of Economics* 14.1, pp. 821–851. DOI: 10.1146/annurev-economics-051520-021409.

Chen, Henian, Patricia Cohen, and Sophie Chen (Apr. 2010). "How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies". *Communications in Statistics - Simulation and Computation* 39.4, pp. 860–864. DOI: 10.1080/03610911003650383.

Cohen, Jack (1988). *Statistical power analysis for the behavioral sciences*. 2nd ed. L. Erlbaum Associates.

Cunningham, Scott (Aug. 2021). *Causal Inference: The Mixtape*. 1st ed. Yale University Press.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt (Oct. 2019). "Predict science to improve science". *Science* 366.6464, pp. 428–429. DOI: 10.1126/science.aaz1704.

Dette, Holger and Martin Schumann (Feb. 2024). "Testing for equivalence of pre-trends in difference-in-differences estimation". *Journal of Business & Economic Statistics* Forthcoming. DOI: 10.1080/07350015.2024.2308121.

Dimmery, Drew (Mar. 2016). *rdd: Regression discontinuity estimation*. DOI: 10.32614/CRAN.package.rdd.

Dreber, Anna, Magnus Johannesson, and Yifan Yang (Mar. 2024). "Selective reporting of placebo tests in top economics journals". *Economic Inquiry* Forthcoming. DOI: 10.1111/ecin.13217.

Eggers, Andrew C. et al. (Jan. 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, pp. 259–274. DOI: 10.1111/ajps.12127.

Fitzgerald, Jack (May 2024). *The need for equivalence testing in economics*. Discussion Paper 125. Institute for Replication. URL: https://www.econstor.eu/handle/10419/296190.

Frandsen, Brigham R. (May 2017). "Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete". *Advances in Econometrics* 38, pp. 281–315. DOI: 10.1108/s0731-905320170000038012.

Gelman, Andrew and Guido Imbens (July 2018). "Why high-order polynomials should not be used in regression discontinuity designs". *Journal of Business & Economic Statistics* 37.3, pp. 447–456. DOI: 10.1080/07350015.2017.1366909.

Gerard, François, Miikka Rokkanen, and Christoph Rothe (July 2020). "Bounds on treatment effects in regression discontinuity designs with a manipulated running variable". *Quantitative Economics* 11.3, pp. 839–870. DOI: 10.3982/qe1079.

Gopalan, Maithreyi, Kelly Rosinger, and Jee Bin Ahn (Mar. 2020). “Use of quasi-experimental research designs in education research: Growth, promise, and challenges”. *Review of Research in Education* 44.1, pp. 218–243. DOI: 10.3102/0091732x20903302.

Hartman, Erin (Oct. 2021). “Equivalence testing for regression discontinuity designs”. *Political Analysis* 29.4, pp. 505–521. DOI: 10.1017/pan.2020.43.

Hartman, Erin and F. Daniel Hidalgo (Oct. 2018). “An equivalence approach to balance and placebo tests”. *American Journal of Political Science* 62.4, pp. 1000–1013. DOI: 10.1111/ajps.12387.

Horrace, William C. and Ronald L. Oaxaca (Mar. 2006). “Results on the bias and inconsistency of ordinary least squares for the linear probability model”. *Economics Letters* 90.3, pp. 321–327. DOI: 10.1016/j.econlet.2005.08.024.

Huntington-Klein, Nick (Aug. 2022). *The Effect: An Introduction to Research Design and Causality*. 1st ed. CRC Press.

Igarashi, Gaku (Nov. 2023). “A nonparametric discontinuity test of density using a beta kernel”. *Journal of Nonparametric Statistics* 35.2, pp. 323–354. DOI: 10.1080/10485252.2022.2150766.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart (Apr. 2008). “Misunderstandings between experimentalists and observationalists about causal inference”. *Journal of the Royal Statistical Society Series A: Statistics in Society* 171.2, pp. 481–502. DOI: 10.1111/j.1467-985x.2007.00527.x.

Imbens, Guido W and Jeffrey M Wooldridge (Mar. 2009). “Recent developments in the econometrics of program evaluation”. *Journal of Economic Literature* 47.1, pp. 5–86. DOI: 10.1257/jel.47.1.5.

Imbens, Guido W. (Dec. 2020). “Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics”. *Journal of Economic Literature* 58.4, pp. 1129–1179. DOI: 10.1257/jel.20191597.

— (Apr. 2024). “Causal inference in the social sciences”. *Annual Review of Statistics and Its Application* 11.1, pp. 123–152. DOI: 10.1146/annurev-statistics-033121-114601.

Imbens, Guido W. and Thomas Lemieux (Feb. 2008). “Regression discontinuity designs: A guide to practice”. *Journal of Econometrics* 142.2, pp. 615–635. DOI: 10.1016/j.jeconom.2007.05.001.

Lakens, Daniël, Anne M. Scheel, and Peder M. Isager (June 2018). “Equivalence testing for psychological research: A tutorial”. *Advances in Methods and Practices in Psychological Science* 1.2, pp. 259–269. DOI: 10.1177/2515245918770963.

Lee, David S and Thomas Lemieux (June 2010). “Regression discontinuity designs in economics”. *Journal of Economic Literature* 48.2, pp. 281–355. DOI: 10.1257/jel.48.2.281.

Ma, Jun, Hugo Jales, and Zhengfei Yu (July 2020). “Minimum contrast empirical likelihood inference of discontinuity in density”. *Journal of Business & Economic Statistics* 38.4, pp. 934–950. DOI: 10.1080/07350015.2019.1617155.

McCrary, Justin (Feb. 2008). “Manipulation of the running variable in the regression discontinuity design: A density test”. *Journal of Econometrics* 142.2, pp. 698–714. DOI: 10.1016/j.jeconom.2007.05.005.

Otsu, Taisuke, Ke-Li Xu, and Yukitoshi Matsushita (Oct. 2013). “Estimation and inference of discontinuity in density”. *Journal of Business & Economic Statistics* 31.4, pp. 507–524. DOI: 10.1080/07350015.2013.818007.

Piaggio, Gilda et al. (Dec. 2012). “Reporting of noninferiority and equivalence randomized trials”. *JAMA* 308.24, pp. 2594–2604. DOI: 10.1001/jama.2012.87802.

Samii, Cyrus (July 2016). “Causal empiricism in quantitative research”. *The Journal of Politics* 78.3, pp. 941–955. DOI: 10.1086/686690.

Schuirmann, Donald J. (Dec. 1987). “A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability”. *Journal of Pharmacokinetics and Biopharmaceutics* 15.6, pp. 657–680. DOI: 10.1007/bf01068419.

Sieweke, Jost and Simone Santoni (Feb. 2020). “Natural experiments in leadership research: An introduction, review, and guidelines”. *The Leadership Quarterly* 31.1, p. 101338. DOI: 10.1016/j.leaqua.2019.101338.

Stanley, T. D., Hristos Doucouliagos, and John P. Ioannidis (July 2022). “Beyond random effects: When small-study findings are more heterogeneous”. *Advances in Methods and Practices in Psychological Science* 5.4, p. 251524592211204. DOI: 10.1177/25152459221120427.

Stanley, T.D. et al. (May 2023). “Unrestricted weighted least squares represent medical research better than random effects in 67,308 Cochrane Meta-analyses”. *Journal of Clinical Epidemiology* 157, pp. 53–58. DOI: 10.1016/j.jclinepi.2023.03.004.

Stommes, Drew, P. M. Aronow, and Fredrik Sävje (Apr. 2023a). “On the reliability of published findings using the regression discontinuity design in political science”. *Research & Politics* 10.2, p. 205316802311664. DOI: 10.1177/20531680231166457.

— (Mar. 2023b). *Replication Data for: On the reliability of published findings using the regression discontinuity design in political science*. Dataset V1. Cambridge, MA, U.S.A.: Harvard Dataverse. DOI: 10.7910/DVN/XT15Y0.

Online Appendix

A Quadratic Solutions and Hartman ECI Intractability

Per Definition 3.4, if $\hat{f}_+(c) < \hat{f}_-(c)$, then computing the Hartman ECI requires solving

$$z_\alpha^* = \frac{\hat{f}_+(c) - \frac{\hat{f}_-(c)}{\epsilon^*}}{\sqrt{\mathrm{Var}\left(\hat{f}_+(c)\right) + \frac{1}{(\epsilon^*)^2}\mathrm{Var}\left(\hat{f}_-(c)\right)}}$$

for ϵ^*. After simplification, this resolves to

$$\left[\mathrm{Var}\left(\hat{f}_-(c)\right)(z_\alpha^*)^2\right](\epsilon^*)^2 + \left[\hat{f}_-(c)\right]\epsilon + \left[\mathrm{Var}\left(\hat{f}_+(c)\right)(z_\alpha^*)^2 - \hat{f}_+(c)\right] = 0. \tag{A1}$$

This equation has a quadratic solution:

$$\epsilon^* = \frac{-\hat{f}_-(c) \pm \sqrt{\left(\hat{f}_-(c)\right)^2 - 4\,(z_\alpha^*)^4\,\mathrm{Var}\left(\hat{f}_-(c)\right)\mathrm{Var}\left(\hat{f}_+(c)\right) + 4\,(z_\alpha^*)^2\,\hat{f}_+(c)\mathrm{Var}\left(\hat{f}_-(c)\right)}}{2\,(z_\alpha^*)^2\,\mathrm{Var}\left(\hat{f}_+(c)\right) - 2\hat{f}_+(c)}. \tag{A2}$$

In contrast, if $\hat{f}_+(c) > \hat{f}_-(c)$, then computing the Hartman ECI requires solving

$$-z_\alpha^* = \frac{\hat{f}_+(c) - \epsilon^*\hat{f}_-(c)}{\sqrt{\mathrm{Var}\left(\hat{f}_+(c)\right) + (\epsilon^*)^2\,\mathrm{Var}\left(\hat{f}_-(c)\right)}}$$

for ϵ^*. After simplification, one obtains

$$\left[\mathrm{Var}\left(\hat{f}_+(c)\right)(z_\alpha^*)^2 - \hat{f}_+(c)\right](\epsilon^*)^2 + \left[\hat{f}_-(c)\right]\epsilon + \left[\mathrm{Var}\left(\hat{f}_-(c)\right)(z_\alpha^*)^2\right] = 0. \tag{A3}$$

This equation also has a quadratic solution:

$$\epsilon^* = \frac{-\hat{f}_-(c) \pm \sqrt{\left(\hat{f}_-(c)\right)^2 - 4\,(z^*_\alpha)^4\,\mathrm{Var}\left(\hat{f}_-(c)\right)\mathrm{Var}\left(\hat{f}_+(c)\right) + 4\,(z^*_\alpha)^2\,\hat{f}_+(c)\mathrm{Var}\left(\hat{f}_-(c)\right)}}{2\,(z^*_\alpha)^2\,\mathrm{Var}\left(\hat{f}_-(c)\right)}. \tag{A4}$$

Notice that the radicands of the quadratic solutions in Equations A2 and A4 are equivalent and equal

$$\left(\hat{f}_-(c)\right)^2 - 4\,(z^*_\alpha)^4\,\mathrm{Var}\left(\hat{f}_-(c)\right)\mathrm{Var}\left(\hat{f}_+(c)\right) + 4\,(z^*_\alpha)^2\,\hat{f}_+(c)\mathrm{Var}\left(\hat{f}_-(c)\right).$$

This arises because for a quadratic equation of the form

$$\epsilon^* = \frac{-v \pm \sqrt{v^2 - 4uw}}{2u},$$

v is equivalent between Equations A1 and A3; specifically, $v = \hat{f}_-(c)$. These two equations also share u and w terms that change positions between equations. Thus the radicands of both Equation A2 and Equation A4 turn negative whenever

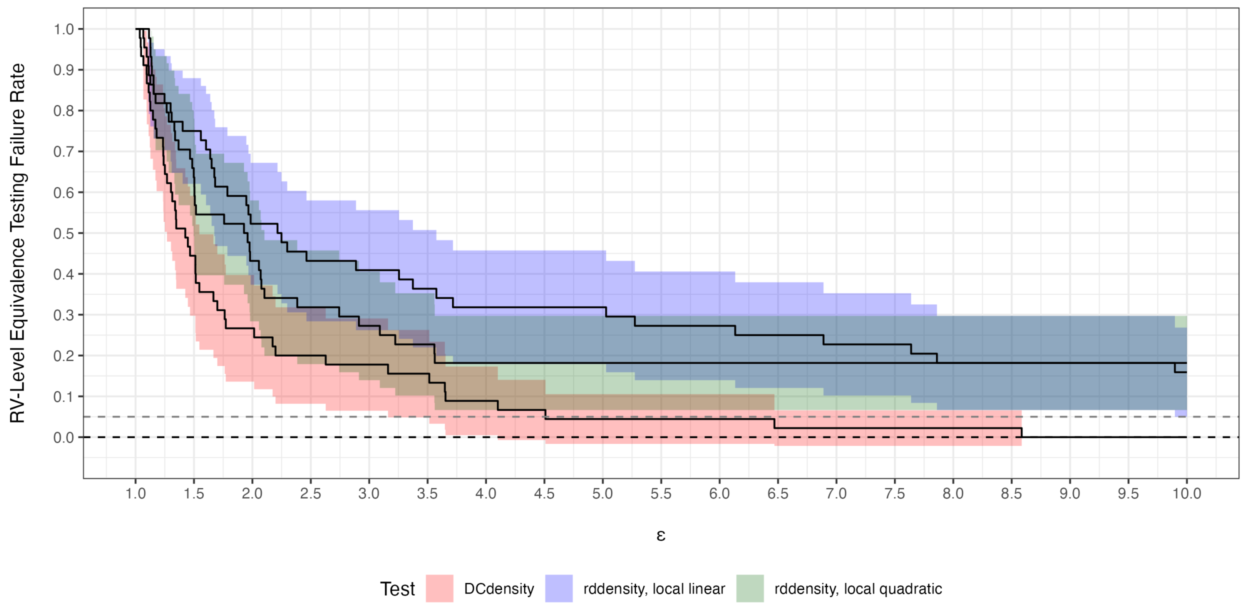
$$\begin{gathered} 4uw > v^2 \\ 4\,(z^*_\alpha)^4\,\mathrm{Var}\left(\hat{f}_-(c)\right)\mathrm{Var}\left(\hat{f}_+(c)\right) > \left(\hat{f}_-(c)\right)^2 + 4\,(z^*_\alpha)^2\,\hat{f}_+(c)\mathrm{Var}\left(\hat{f}_-(c)\right). \end{gathered}$$

B Appendix Tables and Figures

	(1)	(2)	(3)	(4)	(5)	(6)
Failure Rate	0.444	0.75	0.636	0.55	0.571	0.831
	(0.075)	(0.066)	(0.073)	(0.082)	(0.082)	(0.06)
Aggregation Level	RV	RV	RV	Article	Article	Article
Effect Size Measure	DCdensity	rddensity	rddensity	DCdensity	rddensity	rddensity
Estimation Type		Local Linear	Local Quadratic		Local Quadratic	Local Quadratic

Note: This table provides the numerical estimates displayed in Figure 1.

Table A1: Main Failure Rate Estimates



Note: Failure curves are displayed with uncertainty bands representing 95% confidence intervals of RV-level equivalence testing failure rates, based on the standard error of the mean. The black and gray dashed horizontal lines respectively denote 0% and 5% failure rates.

Figure A1: Failure Curves for Different Testing Procedures