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

Jack Fitzgerald, Vrije Universiteit Amsterdam and Tinbergen Institute*

December 17, 2024

Abstract

Researchers applying regression discontinuity design (RDD) often test for endogenous running variable (RV) manipulation around treatment cutoffs, but misinterpret *statistically insignificant* RV manipulation as evidence of *negligible* RV manipulation. I introduce novel procedures that can provide statistically significant evidence that RV manipulation around a cutoff is bounded beneath practically negligible levels. The procedures augment the classic McCrary (2008) density test with an equivalence testing framework, along with bootstrap methods for (cluster-)robust inference. I apply these procedures to replication data from 36 RDD publications, conducting 45 equivalence-based RV manipulation tests. Over 44% of RV density discontinuities at the cutoff cannot be significantly bounded beneath a 50% upward jump. Obtaining equivalence testing failure rates beneath 5% requires arguing that a 350% upward RV density jump at the cutoff is practically equal to zero. My results imply that meaningful RV manipulation around treatment cutoffs cannot be ruled out in many published RDD papers, and that standard tests frequently misclassify the practical significance of RV manipulation. I provide research guidelines and the `lddtest` command in R and Stata to help researchers conduct more credible equivalence-based manipulation testing in future RDD research.

Keywords: `rddensity`, `DCdensity`, cluster bootstrap

JEL: C12, C18, C87, P00

*Email: j.f.fitzgerald@vu.nl.

1 Introduction

Regression discontinuity design (RDD) is one of the cornerstone quasi-experimental techniques that has propelled the credibility revolution in the social sciences over the past 25 years. More than 5600 papers mentioning RDD were published in 2019 alone (Cunningham 2021). 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, eliminating 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. Causal 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. However, in many settings, agents can endogenously manipulate their observed RV values to opt themselves into or out of treatment (Gerard, Rokkanen, & Rothe 2020). Under such RV manipulation, RDD estimates are confounded, reflecting not just causal effects, but also differences in relevant characteristics between agents with different manipulation strategies (see Lee & Lemieux 2010; Gerard, Rokkanen, & Rothe 2020; Cunningham 2021).

RV density discontinuity tests are a popular tool for assessing *ex post* whether such RV manipulation around the cutoff has occurred. McCrary (2008) proposed the first such test, positing 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 over 1750 citations, and that Cattaneo, Jansson, & Ma (2018) and Cattaneo, Jansson, & Ma (2020) already have over 450 citations between them. These tests are a standard recommendation in texts on RDD and causal inference, and are thus functionally required in RDD papers by journal editors and referees in economics, political science, and other disciplines (see Lee & Lemieux 2010; Eggers et al. 2015; Cunningham 2021; Hartman 2021; Huntington-Klein 2022).

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 procedures that are more appropriate for demonstrating that RV manipulation around a cutoff is practically equal to zero. Under these procedures, 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 effectively specifies the largest RV density discontinuity that would be 'economically insignificant'. The pro-

cedures then assess whether there is statistically significant evidence that the right-to-left ratio between these density estimates is significantly bounded beneath this threshold. These approaches build off the widely-used `DCdensity` procedure (McCrary 2008), and augment this procedure with bootstrap methods for finite-sample (cluster-)robust inference.

I demonstrate the necessity of equivalence testing for RV manipulation tests by reanalyzing replication data on 45 RVs used in 36 published RDD articles (see Stommes, Aronow, & Sävje 2023a). 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. My 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, and that such frameworks often misclassify the practical significance of RV manipulation around the cutoff.

Given the clear need for equivalence testing approaches for assessing RV manipulation, I conclude by providing guidelines on credible RV manipulation testing. I 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 given the research setting at hand. I then provide commands in Stata and R that can be used to conduct such testing, including the `lddtest` command in the `eqtesting` R package and the `lddtest` command in Stata; Section 3.3 provides download instructions from Github. These procedures can more credibly rule out practically meaningful RV manipulation around the cutoff in future RDD research.

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 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 estimates 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 $\hat{\theta} \equiv \ln\left(\hat{f}_{+}(c)\right) - \ln\left(\hat{f}_{-}(c)\right)$, and McCrary (2008) provides an asymptotically consistent standard error estimate $\text{SE}\left(\hat{\theta}\right)$; see Sections 3.1 and 3.2 of McCrary (2008) for precise computational details. The procedure concludes by testing whether $\hat{\theta}$ is statistically significantly different from zero, assessing the extremity of test statistic $\frac{\hat{\theta}}{\text{SE}\left(\hat{\theta}\right)}$ 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). This procedure differs from `DCdensity` in at least two important respects. 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 a jackknife estimator

[1] `DCdensity` support is provided in Stata by the `DCdensity.ado` file hosted by Justin McCrary at https://eml.berkeley.edu/∼jmccrary/DCdensity/ (accessed 17 December 2024), and in R *via* the `rdd` R package (Dimmery 2016).

of $\widehat{\text{Var}}\left(\hat{f}_+(c) - \hat{f}_-(c)\right)$, though an alternative asymptotic plug-in variance estimator 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.

In practice, researchers are seldom interested in using these tests to demonstrate that there *is* evidence of RV manipulation around the cutoff. Rather, researchers typically use these tests to demonstrate that such manipulation does *not* occur, and thus that endogenous RV manipulation does not threaten the validity of their causal identification strategy. In practice, researchers 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 and in texts that advocate for the usage of these tests (e.g., see McCrary 2008; Lee & Lemieux 2010; Eggers et al. 2015; Cattaneo, Jansson, & Ma 2018; Cattaneo, Jansson, & Ma 2020; Cunningham 2021; Huntington-Klein 2022).

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

$$\begin{aligned} H_0 &: \lim_{Z \to c^-} f(Z) = \lim_{Z \to c^+} f(Z) \\ H_A &: \lim_{Z \to c^-} f(Z) \neq \lim_{Z \to c^+} f(Z). \end{aligned} \tag{1}$$

These tests are built on the standard null hypothesis significance testing (NHST) framework, as are

other proposed alternative RV manipulation tests (e.g., see Otsu, Xu, & Matsushita 2013; Frandsen 2017; Bugni & Canay 2021; Ma, Jales, & Yu 2021; Igarashi 2023). Researchers typically use statistically insignificant test statistics in RV manipulation tests under standard NHST as evidence in favor of Equation 1's H_0. However, this inference is a well-known misinterpretation of statistical significance (see Altman & Bland 1995; Wasserstein & Lazar 2016).

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. 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). Thus 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 reviewers and 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

$$\begin{aligned} H_0 &: \lim_{Z \to c^-} f(Z) \not\approx \lim_{Z \to c^+} f(Z) \\ H_A &: \lim_{Z \to c^-} f(Z) \approx \lim_{Z \to c^+} f(Z). \end{aligned} \tag{2}$$

If one can set a range within 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 several implementations.

All of the procedures that I describe 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, then 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, then 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 units. 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. This provides useful inference guarantees for tests based on standard Gaussian inference.

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 2 with 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)**.** *One wishes to assess the hypotheses in Equation 2 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 can conclude that $\lim_{Z \to c^{-}} f(Z)$ *is practically equal to* $\lim_{Z \to c^{+}} f(Z)$ *if and only if* $ECI_{1-\alpha} \subset [-\ln(\epsilon), \ln(\epsilon)]$.

Passing the LDD ECI through the exponential function provides the smallest range of ratios

within which one can significantly bound the ratio of $\hat{f}_+(c)$ to $\hat{f}_-(c)$ using the LDD equivalence test described in Definition 3.1. 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 (Cluster) Bootstrap Inference Augmentations

Though `DCdensity` is useful for estimating logarithmic density discontinuities, two inference issues arise in practice. First, McCrary (2008) relies on asymptotic convergence to attain a standard normal distribution for $\hat{\theta}$. However, in practice, RDD is often estimated on a relatively small number of observations, especially when considering effective sample sizes after specifying bandwidths. As I show in Section 4, `DCdensity`'s automatic bandwidth selection procedure decreases effective sample sizes in the median density discontinuity estimation by 34%. Whereas `rddensity` addresses this using jackknife variance estimators by default, `DCdensity` lacks similar resampling approaches for finite-sample robust inference. Second, both `DCdensity` and `rddensity` (implictly) assume that observations are identically and independently distributed. However, RDD applications with clustered data (e.g., panel data) arise in many empirical applications (see Lee & Lemieux 2010). Neither `DCdensity` nor `rddensity` employ methods for cluster-robust inference.

To address these issues, I augment the `DCdensity` procedure with a (cluster) bootstrap approach for (cluster-)robust inference. I provide this procedure in Algorithm 1. This procedure effectively extends the percentile bootstrap advocated by Hanh & Liao (2021) to the equivalence testing context. A p-value for this method can be effectively computed as the proportion of the bootstrap estimates $\hat{\theta}_R$ that fall in the range $[-\ln(\epsilon), \ln(\epsilon)]$.

Algorithm 1 (Cluster) Bootstrapping for Logarithmic Density Discontinuity Equivalence Tests

Require: $\alpha \in (0, 0.5)$; $R > 0$; $\epsilon > 1$; $N > 0$ observations (which may be clustered in $C > 0$ clusters) with defined values of Z.

Estimate $\hat{\theta}$ using `DCdensity` on the full dataset; save bin size b and bandwidth h from `DCdensity`'s automatic selection procedures.

while $R > 0$ **do**

if Data is clustered **then**

Randomly resample C clusters from the dataset with replacement.

else

Randomly resample N observations from the dataset with replacement.

end if

Estimate $\hat{\theta}_R$ using `DCdensity` on the resampled dataset, holding b and h constant.

if $\hat{\theta}_R$ is defined **then**

Store estimate $\hat{\theta}_R$.

$R \leftarrow R - 1$

end if

end while

Order estimates $\hat{\theta}_R$ from least to greatest. Define $\text{ECI}_{1-\alpha}$ as $\left[\hat{\theta}_\alpha, \hat{\theta}_{1-\alpha}\right]$, where $\hat{\theta}_\alpha$ is the α-quantile of the sample of bootstrap estimates $\hat{\theta}_R$. Conclude that $\lim_{Z \to c^-} f(Z)$ is practically equal to $\lim_{Z \to c^+} f(Z)$ if and only if $\text{ECI}_{1-\alpha} \subset [-\ln(\epsilon), \ln(\epsilon)]$.

3.3 Statistical Software Commands

I provide statistical software commands in Stata and R to implement my proposed LDD equivalence testing procedures from Sections 3.1 and 3.2. For Stata, I provide the `lddtest` command, which can be accessed from https://github.com/jack-fitzgerald/lddtest.[2] 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.

3.4 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'.[3]

Definition 3.3 (The Hartman Test)**.** *The researcher wishes to assess the hypotheses in Equation 2 using a test with Type I error rate $\alpha \in (0, 0.5)$. They thus set a maximally acceptable right-to-left*

[2]Bootstrap routines are currently only available for the R version of `lddtest`; such routines for the Stata version are in development.

[3]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 generated by the `rddensity` command in R (Cattaneo, Jansson, & Ma 2020). See Hartman (2021) for details.

density ratio $\epsilon > 1$ *and formulate null and alternative hypotheses*

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

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

$$t_H^- = \frac{\hat{f}_+(c) - \frac{\hat{f}_-(c)}{\epsilon}}{\sqrt{\widehat{Var}\left(\hat{f}_+(c)\right) + \frac{1}{\epsilon^2}\widehat{Var}\left(\hat{f}_-(c)\right)}} \qquad t_H^+ = \frac{\hat{f}_+(c) - \epsilon\hat{f}_-(c)}{\sqrt{\widehat{Var}\left(\hat{f}_+(c)\right) + \epsilon^2\widehat{Var}\left(\hat{f}_-(c)\right)}}, \tag{3}$$

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* $\lim\limits_{Z \to c^-} f(Z)$ *is practically equal to* $\lim\limits_{Z \to c^+} f(Z)$ *if and only if* $t_H \geq z_\alpha^*$. *If* $t_H = t_H^+$, *then the researcher rejects* H_0 *and concludes that* $\lim\limits_{Z \to c^-} f(Z)$ *is practically equal to* $\lim\limits_{Z \to c^+} f(Z)$ *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).[4]

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

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

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

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

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

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 3, 4, and 5 are by default computed using the jackknife variance estimator proposed for `rddensity` by Cattaneo, Jansson, & Ma (2020). At time of writing, `rddensity` does not offer support for cluster-robust inference.

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, considerably skewing estimates (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.

My procedures address this issue by using `DCdensity` rather than `rddensity`. Because `DCdensity`'s $\hat{\theta}$ estimand is the difference between $\ln\left(\hat{f}_+(c)\right)$ and $\ln\left(\hat{f}_-(c)\right)$, $\hat{f}_-(c)$ and $\hat{f}_+(c)$ are both guaranteed to be positive whenever $\hat{\theta}$ is defined. 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 4 or Equation 5 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 solution candidates are negative whenever

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

If this occurs, then no analytic solution for ϵ^{*} exists on the real plane, as solving Equation 4 or Equation 5 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. This implies that one can always find a critical ϵ^{*} for which $\hat{\theta}$ is significantly bounded within $\left[-\ln\left(\epsilon^{*}\right), \ln\left(\epsilon^{*}\right)\right]$.

Consequently, though no valid Hartman ECIs can be defined for several RV density discontinuities in my replication sample, similar issues do not arise for ECIs arising from my proposed equivalence testing procedures. As I discuss in Section 4, when RV density discontinuities are computed using local linear (quadratic) regression under `rddensity`, four (seven) of the 45 RVs in my replication sample have non-tractable Hartman ECIs. However, these tractability issues are completely eliminated for all RVs in my replication sample after switching to testing procedures that employ `DCdensity`.

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.[5] 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.

Though my results arise from data on political science publications, my findings are also 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

[5]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.

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, 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 my `lddtest` command, which computes results from `DCdensity` and applies the bootstrap procedures described in Algorithm 1.[6] 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. Standard errors for `rddensity` are computed using the default jackknife estimator. Though `DCdensity` produces a direct $\hat{\theta}$ point estimate, I can compute an equivalent $\hat{\theta}$ estimate for `rddensity` from its $\hat{f}_{-}(c)$ and $\hat{f}_{+}(c)$ outputs using the identity $\hat{\theta} \equiv \ln\left(\hat{f}_{+}(c)\right) - \ln\left(\hat{f}_{-}(c)\right)$, provided that `rddensity` produces non-negative $\hat{f}_{-}(c)$ and $\hat{f}_{+}(c)$. I drop any `rddensity` esti-

[6] When observations are clustered, I apply the cluster bootstrap routine from Algorithm 1. Otherwise, I simply apply the standard bootstrap routine. Bootstrap results arise from 10,000 bootstrap replications on each RV.

	Min	P10	P25	P50	P75	P90	Max	Mean	SD	N
N	134	257.6	706	1450	21773	114514.6	517255	41070.333	102542.26	45
Effective N, `DCdensity`	90	165.4	302	1183	11633	59022.6	374632	28714.533	79282.174	45
Effective N, `rddensity`, local linear	47	72.8	138	323	1267	13037.6	51198	4139.756	9930.422	45
Effective N, `rddensity`, local quadratic	48	146	255	941	5946	24687.6	81713	7845.067	16271.616	45
% of sample in bandwidth, `DCdensity`	0.225	0.358	0.515	0.66	0.737	0.853	0.94	0.628	0.178	45
% of sample in bandwidth, `rddensity`, local linear	0.001	0.081	0.103	0.173	0.256	0.332	0.766	0.203	0.143	45
% of sample in bandwidth, `rddensity`, local quadratic	0.068	0.138	0.207	0.436	0.648	0.723	0.91	0.449	0.234	45
$\hat{\theta}$, `DCdensity`	-1.141	-0.273	-0.065	0.006	0.237	0.616	1.466	0.072	0.44	45
$\hat{\theta}$, `rddensity`, local linear	-2.223	-0.445	-0.103	0.023	0.474	0.785	1.389	0.085	0.576	44
$\hat{\theta}$, `rddensity`, local quadratic	-2.67	-0.308	-0.106	-0.015	0.092	0.514	1.848	-0.001	0.567	44
$\lvert\hat{\theta}\rvert$, `DCdensity`	0	0.02	0.044	0.14	0.375	0.784	1.466	0.28	0.344	45
$\lvert\hat{\theta}\rvert$, `rddensity`, local linear	0	0.033	0.064	0.235	0.529	0.822	2.223	0.39	0.427	44
$\lvert\hat{\theta}\rvert$, `rddensity`, local quadratic	0	0.024	0.06	0.103	0.315	0.608	2.67	0.291	0.485	44
Standard NHST p-value, `DCdensity`	0	0	0.009	0.187	0.471	0.824	0.999	0.301	0.332	45
Standard NHST p-value, `rddensity`, local linear	0	0.01	0.102	0.364	0.519	0.791	1	0.367	0.296	44
Standard NHST p-value, `rddensity`, local quadratic	0	0.021	0.162	0.558	0.735	0.932	1	0.5	0.322	44
Asymptotic ϵ^*, `DCdensity`	1.038	1.085	1.179	1.426	2.015	3.651	8.586	2.002	1.511	45
Bootstrap ϵ^*, `DCdensity`	1.039	1.084	1.222	1.427	2.008	4.318	16.133	2.456	2.92	45
ϵ^*, `rddensity`, local linear	1.066	1.113	1.309	1.984	5.028	9.896	45.882	4.887	7.793	41
ϵ^*, `rddensity`, local quadratic	1.114	1.136	1.27	1.511	2.096	3.323	25.39	2.717	4.224	38
Asymptotic SE $\left(\hat{\theta}\right)$, `DCdensity`	0.005	0.017	0.04	0.105	0.221	0.369	0.551	0.152	0.142	45
Bootstrap SE $\left(\hat{\theta}\right)$, `DCdensity`	0.012	0.019	0.042	0.102	0.233	0.429	0.935	0.182	0.202	45

Note: Summary statistics are computed across the 45 density discontinuities estimated in my replication sample. The N column denotes the number of RV density discontinuities for which the variable is non-missing.

Table 1: Summary Statistics of Replication Results

mates 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 most variables concerning `rddensity`. This is because both local linear and local quadratic estimates from `rddensity` yield non-positive estimates of $\hat{f}_-(c)$ and/or $\hat{f}_+(c)$ for one RV in the sample.

Though sample sizes for the RVs themselves appear to be reasonably large, effective sample sizes of RV values which lie within each procedure's default bandwidths can be particularly small. The median sample size N is 1450, but `DCdensity`'s default bandwidths reduce effective sample sizes in the median RV by 34%, and `rddensity`'s reduce such median effective sample sizes by 56.4-82.7%. This implies that a considerable proportion of RV density discontinuity estimates may be quite underpowered, particularly for `rddensity`'s relatively strict default bandwidths.

Examining the observation counts for ϵ^* across each of the testing frameworks makes the Hartman ECI's tractability issues apparent. As discussed in Section 3.4, 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 $\hat{f}_{+}(c)$.

As a benchmark, I conduct equivalence testing for each RV density discontinuity by assessing whether there is statistically significant evidence (at $\alpha = 0.05$) 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 asymptotic results from `DCdensity`, the bootstrap routine detailed in Algorithm 1 for bootstrap results from `lddtest`, and the Hartman test described in Definition 3.3 for results from `rddensity`. In the terminology of these tests, I effectively set the threshold $\epsilon = 1.5$.

I select $\epsilon = 1.5$ as my benchmark right-to-left density ratio for three reasons. First, Chen, Cohen, & Chen (2010) provide 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 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.

In RDD publications from top journals, it should be easy to show that RV density discontinuities at the cutoff can be significantly bounded beneath a 50% upward jump. If this condition

	(1)	(2)	(3)	(4)
Equivalence Testing Failure Rate	0.444	0.467	0.75	0.636
	(0.075)	(0.075)	(0.066)	(0.073)
N	45	45	44	44
Procedure	DCdensity	lddtest	rddensity	rddensity
Estimation Type	Asymptotic	Bootstrap	Local Linear	Local Quadratic

Note: Equivalence testing failure rates represent the proportion of logarithmic right-to-left RV density ratios at the cutoff that cannot be significantly bounded within $[-\ln(1.5), \ln(1.5)]$ using the estimation/testing procedure specified by the column. Standard errors of the mean are presented in parentheses.

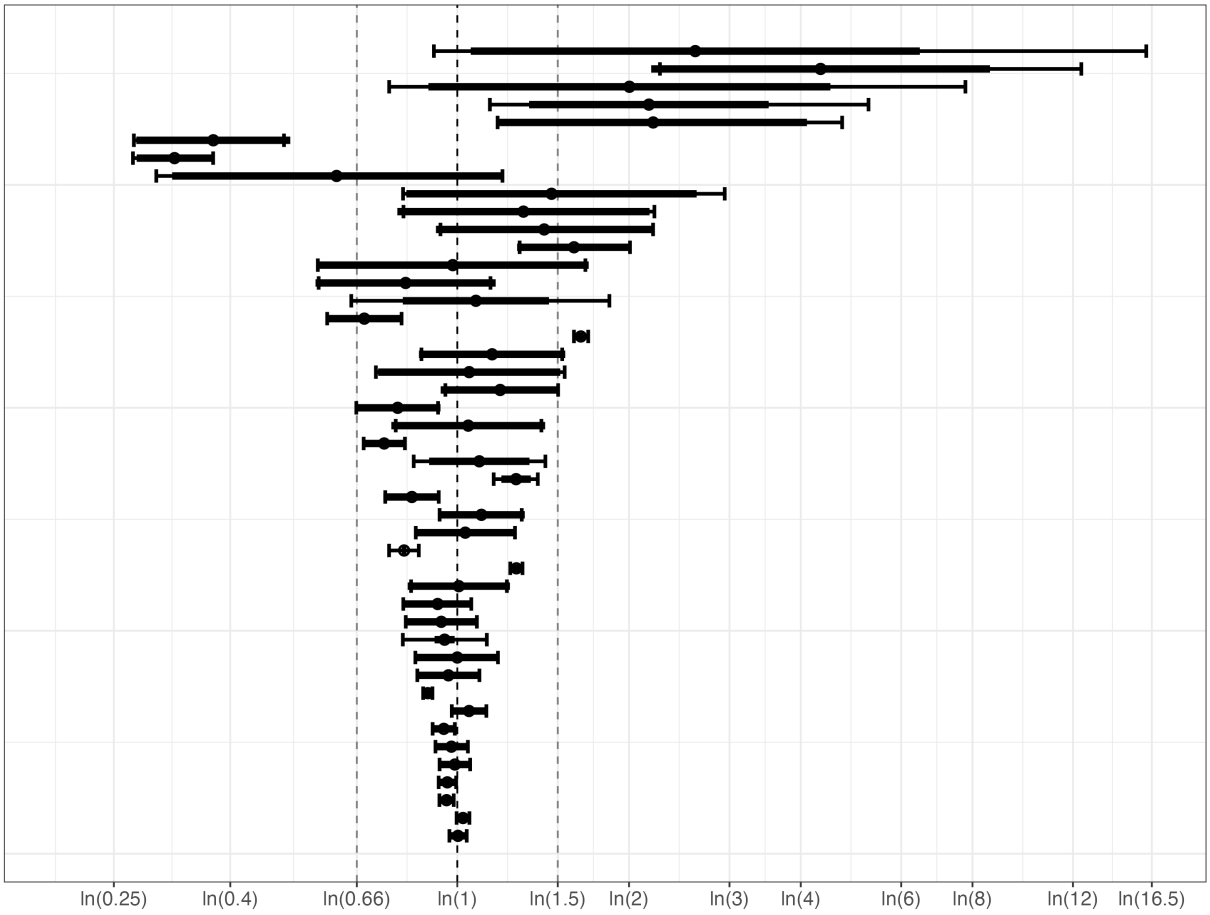
Table 2: Main Equivalence Testing Failure Rate Estimates

holds for a given RV, then this RV 'passes' my lenient benchmark equivalence tests. I compute the proportion of RVs that 'fail' these benchmark equivalence tests, which I term the 'equivalence testing failure rate' (see also Fitzgerald 2024).

5 Results

In my sample, equivalence testing failure rates for RV density discontinuities at the cutoff are quite high. Figure 1 displays the main results from `DCdensity`/`lddtest`, and Table 2 provides equivalence testing failure rates across all tests. When looking to the asymptotic SEs computed by `DCdensity`, 44.4% of the RV density discontinuities at the cutoff cannot be significantly bounded beneath a 50% upward jump. When instead employing the bootstrap approach in `lddtest`, this rate increases to 46.7%. Equivalence testing failure rates are much higher for `rddensity` results examined by the Hartman test, ranging from 63.6% for local quadratic estimates to 75% for local linear estimates.

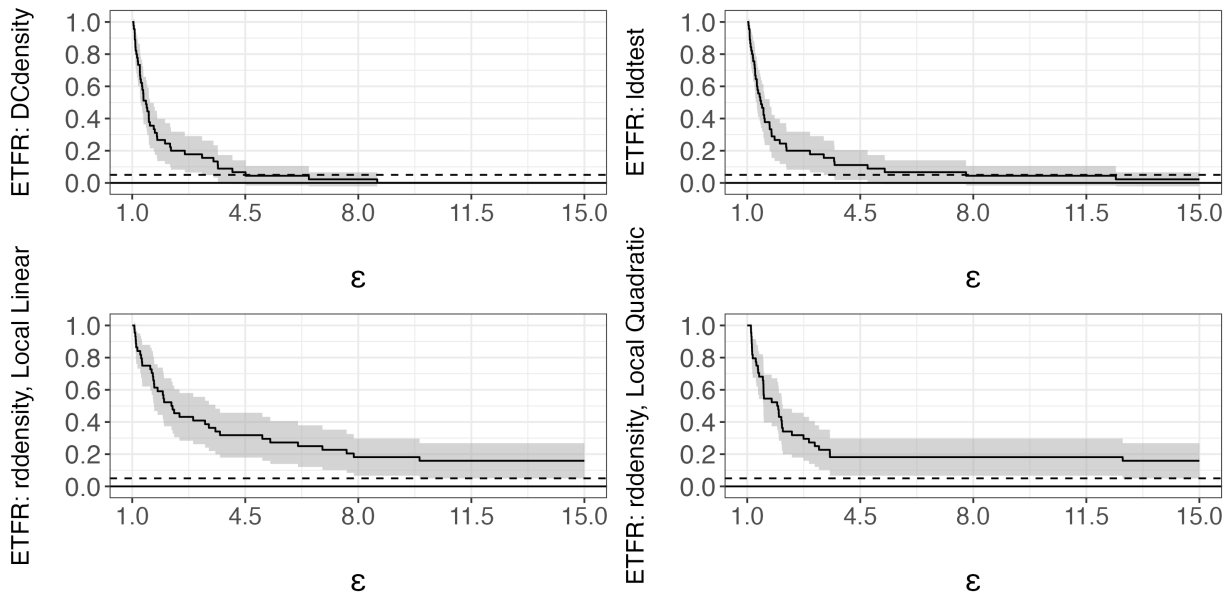
These high equivalence testing failure rates arise not just because RV density discontinuity estimates are underpowered and imprecise, but also because the point estimates themselves are quite large. Online Appendix B details an unrestricted weighted least squares meta-analysis of



Note: Logarithmic RV density discontinuities at the cutoff estimated by `DCdensity/lddtest` are displayed for each of the 45 RVs in my sample, along with asymptotic 95% ECIs estimated using `DCdensity` (thicker bands) and bootstrap 95% ECIs estimated using `lddtest` (thinner bands with error bars). Dashed vertical black and gray lines respectively denote zero and $\pm \ln(1.5)$.

Figure 1: Logarithmic Density Discontinuities Estimated by `DCdensity/lddtest`

absolute logarithmic density discontinuities, which shows that the average RV density discontinuity at the cutoff is equivalent to a 26.5% upward jump (see Stanley et al. 2023). These results imply that treatment effects estimated by RDD setups 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.



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

Figure 2: Failure Curves for Different Testing Procedures

5.1 Failure Curves

My high equivalence testing failure rate estimates cannot be explained by my choice of maximal acceptable right-to-left density ratio ϵ. Figure 2 displays 'failure curves', which show the distribution of equivalence testing failure rates across different ϵ thresholds and testing procedures (see also Fitzgerald 2024). The shapes of the failure curves reflect the fact that equivalence testing failure rates decline when larger RV density discontinuities at the cutoff are tolerated.

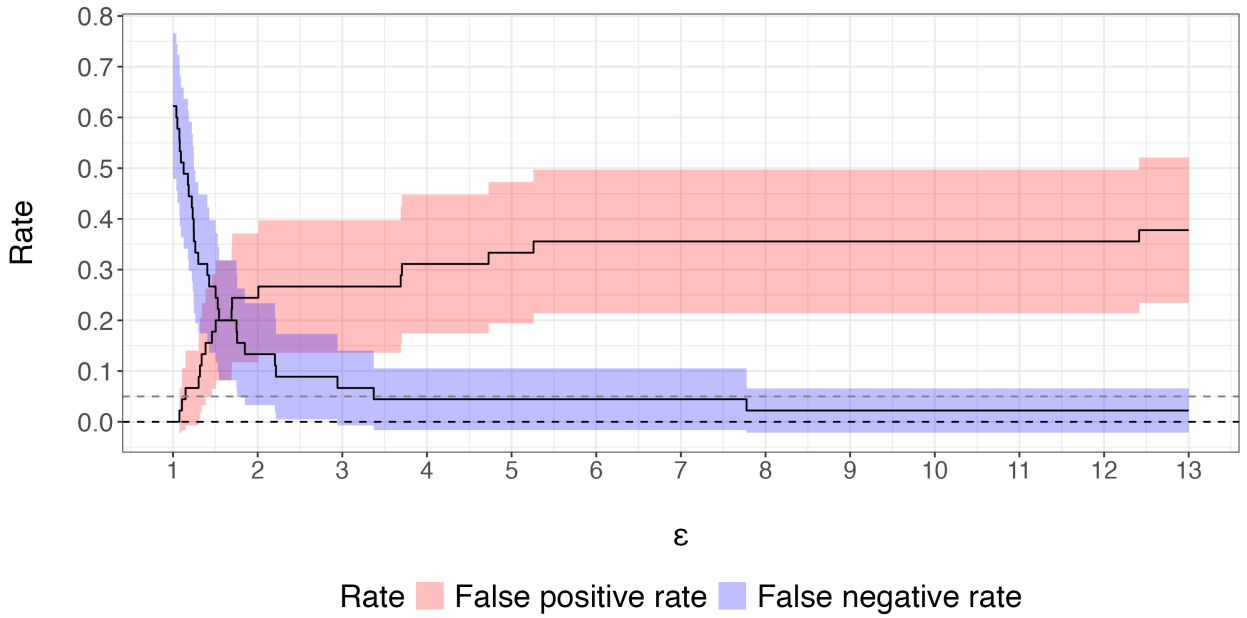
The failure curves in Figure 2 show that in my sample, equivalence testing failure rates for RV manipulation tests remain significantly above nominal levels even as the ϵ threshold is allowed to grow exceedingly large. Consider the RV density discontinuity that one would need to tolerate in order to obtain equivalence testing failure rates beneath a traditional 5% rate. The horizontal dashed lines in Figure 2 denote 5% equivalence testing failure rates; the failure curve for `DCdensity` only crosses this line when $\epsilon = 4.509$. This implies that in order to obtain equivalence testing

failure rates beneath 5%, one must 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. The failure curve for `lddtest` only crosses 5% for $\epsilon = 12.414$, implying that one must be willing to argue that even larger upward jumps are practically negligible to sufficiently bound error rates for other tests. Because such arguments are 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.

5.2 Confusion Curves

As discussed in Section 2, standard NHST can yield misleading results concerning whether RV density discontinuities are well-bounded in two ways. First, false positives can arise when $\hat{\theta}$ estimates are statistically significantly different from zero, yet also significantly bounded beneath practically negligible $\ln(\epsilon)$ thresholds. Second, false negatives can arise when $\hat{\theta}$ is not statistically significantly different from zero, but is also not significantly bounded beneath practically practically negligible ϵ thresholds.

In this context, false positive and false negative rates depend on the ϵ threshold denoting the practically negligible right-to-left density ratio for a given research context. Figure 3 displays 'confusion curves', which show how these false positive and false negative rates are distributed in my replication sample for different values of ϵ. These confusion *curves* extend the well-known concept of a confusion *matrix* to a setting where ground truths depend on a continuous parameter, and are based on replication results that apply the `lddtest` command's bootstrap procedure (see Algo-



Note: Failure curves are displayed with uncertainty bands representing 95% confidence intervals of RV-level false positive and false negative rates, based on the standard errors of the means for bootstrap results from `lddtest`. The black and gray dashed horizontal lines respectively denote 0% and 5% error rates.

Figure 3: Confusion Curves

rithm 1). For a given ϵ, false positives arise if over 95% of the bootstrap estimates $\hat{\theta}_R$ are on the same side of zero as the $\hat{\theta}$ point estimate, while simultaneously over 90% of the $\hat{\theta}_R$ estimates are bounded in $[-\ln(\epsilon), \ln(\epsilon)]$. Likewise, false negatives arise if no more than 95% of the $\hat{\theta}_R$ estimates are on the same side of zero as $\hat{\theta}$, while simultaneously no more than 90% of the $\hat{\theta}_R$ estimates are bounded in $[-\ln(\epsilon), \ln(\epsilon)]$.

The confusion curves in Figure 3 show that typical RV manipulation tests based on standard NHST frequently misclassify the practical significance of RV density discontinuity estimates. At the benchmark effect size of $\epsilon = 1.5$, 17.7% of the RV density discontinuity estimates in my sample are 'false positives' that are statistically significantly different from $\epsilon = 1$, yet significantly bounded beneath $\epsilon = 1.5$. These estimates would be flagged as evidence of problematic RV manipulation by standard testing frameworks despite being significantly bounded beneath practically negligible levels (provided that $\epsilon = 1.5$ is a good threshold for practically negligible $\hat{\theta}$ estimates

given the research setting at hand). Perhaps more worryingly, at $\epsilon = 1.5$, 26.6% of RV density discontinuity estimates in my sample are 'false negatives' that are neither statistically significantly different from $\epsilon = 1$ nor significantly bounded beneath $\epsilon = 1.5$ These estimates would 'pass' standard RV manipulation tests despite there being no statistically significant evidence that these RV density discontinuities can be bounded beneath practically negligible levels. When applying standard RV manipulation tests, estimates of the former sort may result in perfectly valid RVs being discarded for research simply because their density discontinuities at the cutoff are precisely estimated. Conversely, when applying standard RV manipulation tests, estimates of the latter sort may mislead researchers into believing that meaningfully manipulated RVs are suitable for RDD simply because their density discontinuities at the cutoff are imprecisely estimated.

6 Conclusion

I introduce several equivalence-based RV manipulation tests for RDD applications. In a large sample of RDD publications in top 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. 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.

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. 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 pro-

posed equivalence testing approaches 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 3.4 and Section 4). My recommended procedures 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 3.3 provides download instructions.

In the event that my testing procedures do not yield statistically significant evidence that RV manipulation near the cutoff is negligible, I recommend that researchers employ alternative procedures to assess the robustness of their estimates to RV manipulation. Specifically, in this case, I recommend that researchers use the `rdbounds` procedure in Stata and R developed by Gerard, Rokkanen, & Rothe (2020), which partially identifies local average treatment effects in RDD settings with RV manipulation at the cutoff. Because significant evidence of negligible RV manipulation is neither strictly necessary nor sufficient to ensure that research conclusions arising from RDD are robust to RV manipulation, it may also be reasonable for researchers to abandon RV manipulation testing entirely, instead simply assessing whether their original RDD estimates yield the same research conclusions as those arising from the `rdbounds` procedure. However, if RV manipulation tests are to be conducted, then the equivalence testing approaches that I propose in this paper provide more credible evidence than existing tests that RDD is the natural experiment that it promises to be.

7 Acknowledgments

I thank Peter Hull, Michal Kolesár, Didier Nibbering, conference participants from the Causal Data Science Meeting and the University of Groningen Workshop on Causal Inference + Machine Learning, as well as seminar participants from Vrije Universiteit Amsterdam for helpful comments and feedback. All errors are my own.

8 Declaration of Interest

I am grateful to the Amsterdam Law and Behavior Institute for PhD funding. Writing on this manuscript began when I was serving 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 Ethical Review Board of Vrije Universiteit Amsterdam's School of Business and Economics.

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.

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.

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.

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.

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*. Institute for Replication Discussion Paper Series No. 125. 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.

Hahn, Jinyong and Zhipeng Liao (2021). "Bootstrap standard error estimates and inference". *Econometrica* 89.4, pp. 1963–1977. DOI: 10.3982/ecta17912.

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

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.

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.

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.

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

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

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

for ϵ^*. After simplification, this resolves to

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

This equation has a quadratic solution:

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

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

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

for ϵ^*. After simplification, one obtains

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

This equation also has a quadratic solution:

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

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

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

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

$$4uw > v^2$$

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

B Meta-Analysis

How large is the average RV density jump at the cutoff? To examine this question, I obtain meta-analytic estimates of *absolute* logarithmic RV density discontinuities. 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 estimates using an unrestricted weighted least squares approach (Stanley et al. 2023). Let i index a given RV. This approach employs a regression of the form

$$\frac{\left|\hat{\theta}_i\right|}{\text{SE}\left(\hat{\theta}_i\right)} = \beta \frac{1}{\text{SE}\left(\hat{\theta}_i\right)} + \mu_i, \tag{A5}$$

where β is the meta-analytic estimate of interest and $\text{SE}\left(\hat{\theta}_i\right)$ is parametrically estimated via `DCdensity`. In this setting, β is equivalent to a weighted average of effect sizes $\left|\hat{\theta}_i\right|$, where the weights are given by $\left(\text{SE}\left(\hat{\theta}_i\right)\right)^{-2}$ (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).

The meta-analytic estimate for $\left|\hat{\theta}\right| = 0.235$, with a standard error $\text{SE}\left(\left|\hat{\theta}\right|\right) = 0.025$. This estimate implies that the meta-analytic average RV density discontinuity at the cutoff is equivalent to a 26.5% upward jump. This average discontinuity is precisely estimated, and is quite significantly bounded beneath a threshold of $\ln(1.5)$. However, 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 this magnitude.