PA

Equivalence Testing for Regression Discontinuity Designs

Erin Hartman

Assistant Professor of Political Science and Statistics, University of California, Los Angeles, CA, USA.
Email: ekhartman@ucla.edu, URL: www.erinhartman.com

Abstract

Regression discontinuity (RD) designs are increasingly common in political science. They have many advantages, including a known and observable treatment assignment mechanism. The literature has emphasized the need for "falsification tests" and ways to assess the validity of the design. When implementing RD designs, researchers typically rely on two falsification tests, based on empirically testable implications of the identifying assumptions, to argue the design is credible. These tests, one for continuity in the regression function for a pretreatment covariate, and one for continuity in the density of the forcing variable, use a null of no difference in the parameter of interest at the discontinuity. Common practice can, incorrectly, conflate a failure to reject evidence of a flawed design with evidence that the design is credible. The well-known equivalence testing approach addresses these problems, but how to implement equivalence tests in the RD framework is not straightforward. This paper develops two equivalence tests tailored for RD designs that allow researchers to provide statistical evidence that the design is credible. Simulation studies show the superior performance of equivalence-based tests over tests-of-difference, as used in current practice. The tests are applied to the close elections RD data presented in Eggers et al. (2015b).

Keywords: regression discontinuity design, falsification tests, equivalence tests

1 Introduction

The regression discontinuity (RD) design is an observational causal identification strategy used to study the impact of a deterministic treatment assignment mechanism, such as the incumbency effect for a party that wins a close election. Treatment is assigned based on the value of a score, referred to as the forcing variable, such that all units with a score below a cutoff do not receive treatment, and all the units with a score above the cutoff do. The method was first described by Thistlethwaite and Campbell (1960) with statistical properties derived in Hahn, Todd, and van der Klaauw (2001) and Lee (2008); they have many advantages and have become increasingly popular in political science (Skovron and Titiunik 2015), in part because they have known, observable treatment assignment mechanisms (Cattaneo, Idrobo, and Titiunik 2020). RD designs are typically thought to require relatively weak assumptions compared to other common analysis techniques for observational studies, such as regression or instrumental variables (De la Cuesta and Imai 2016).

While the assumptions may be weaker than some observational methods, RD designs still rely on strong causal identification assumptions. As Eggers *et al*. (2015b, p. 270) state, "the burden of proof is on the researcher to justify her assumptions and subject them to rigorous testing." While the necessary assumptions cannot be directly empirically tested, the literature suggests that researchers should (1) consider theoretical mechanisms under which RD designs could be invalidated and (2) use falsification tests, that is, statistical hypothesis tests of observable implications of the necessary assumptions, to bolster their claims that the RD design is credible (Eggers *et al*. 2015b). Based on the suggestion of Hartman and Hidalgo (2018), I argue that in order to place the burden of proof on the researcher, and allow the data to support the design, falsification tests

Political Analysis (2020)

DOI: 10.1017/pan.2020.43

Corresponding author
Erin Hartman

Edited by
Jeff Gill

should be structured such that:

$$\begin{aligned} H_0 &: \text{The data are } \textit{inconsistent} \text{ with observable implications} \\ &\quad \text{of a valid regression discontinuity design} \\ H_A &: \text{The data are } \textit{consistent} \text{ with observable implications} \\ &\quad \text{of a valid regression discontinuity design.} \end{aligned} \tag{1}$$

Based on the assumptions outlined in Lee (2008), there are two common statistical tests, derived from observable implications of the identifying assumptions, used in evaluating RD designs—(1) that the limits for the regression function of *any* pretreatment covariate is continuous at the cutoff (often referred to as a "balance" or "placebo" test) and (2) that the density of the forcing variable is continuous at the cutoff. In current practice, these tests are usually structured using a test-of-difference in which H_0 is that the difference in the limits is zero.[1] This formulation, which is a common practice across many observational causal designs, is reversed from the structure I suggest above in Equation (1).

By focusing on failure to reject the null in a test-of-difference, researchers reverse the appropriate type I error necessary to support a credible design; they should use a test that controls the probability of falsely rejecting evidence of a *flawed* design. While merely flipping the null and alternative hypothesis might seem like a small matter, the conflation of statistical insignificance with evidence in favor of a credible design is problematic in many ways. RD designs often have relatively little data near the cutoff, leading to little power to reject a null of no difference at the cutoff. This lack of power can lead a researcher to erroneously disregard a large point estimate that is statistically insignificant, even if there is, in truth, not enough power to detect such a discontinuity at the cutoff.[2] For example, Eggers *et al*. (2015b) conduct placebo tests for whether a party won the previous election and previous vote margin. The researchers present *p*-values for different close-election settings, but do not present point estimates. In the reanalysis presented in Figure 1, I find that 8 of the 24 point estimates are larger than 2.5 percentage points. These substantively large estimates are masked by relying solely on the *p*-value of the test.

Conversely, more data, which leads to increased power, can lead researchers to conclude that a small point estimate, which is trivially close to, but not exactly, zero is indicative of a flawed design, even if such a difference is substantively inconsequential. Researchers acknowledge this point, and will focus on the substantive size of point estimates if they are statistically significant, however they do not necessarily discuss the substantive size of the estimate if they fail to reject a null of equality. Using a standard null hypothesis test for falsification testing is not problematic, per se, since a large *p*-value does not imply that imbalance is inconsequential. As is clear, though, if these tests are used to evaluate the credibility of the design, avoiding conflation of statistical power and substantive equivalence can require ad hoc justification of results by researchers. A key advantage of the equivalence approach is that judgments about what counts as a substantively large or small imbalance is made more explicit and the resulting analysis is more transparent.

1 Cattaneo *et al*. (2020) discuss three additional tests: treatment effects at artificial cutoffs, exclusion of observations near the cutoffs, and sensitivity to the selection of bandwidth. These tests all implement variants of the statistical tests described in this manuscript.

2 Researchers acknowledge that failure to reject the null in a test-of-difference is not evidence of equivalence, and some researchers therefore recommend conducting the test at a larger α level of 0.15 or 0.2. For example, Cattaneo *et al*. (2015, p. 9) say "As our focus is on type II error, [the α] value should be chosen to be higher than conventional levels for a conservative choice for W_0 [the window size]. Based on the power calculations discussed above, a reasonable choice is to adopt $\alpha = 0.15$; higher values will lead to a more conservative choice of W_0 if a feasible window satisfies the stricter requirement." While this purpose-specific approach can address the power concerns, the advantage of the equivalence-based approach is that the test will control the type I error rate as defined by α without the need for the researcher to do adjustment to address power. Construction of 100(1- α)% equivalence confidence intervals is also straightforward.

By restructuring the hypotheses as in Equation (1), researchers can provide statistical evidence that supports the credibility of the design, rather than merely failing to detect evidence of a flawed design. I provide a set of equivalence tests—statistical tests with a null hypothesis that the parameter of interest is outside of a "substantively inconsequential" range, defined by the researcher, versus an alternative that it is within this range—for use in evaluating RD designs. A skeptical researcher will only reject the null of an important difference in favor of a substantively inconsequential difference with sufficient support from the data. How to apply these equivalence tests in the RD setting is not straightforward, given the unique estimators required in RD to estimate the effect at the cutoff, as well as the need to define "equivalence" for a density function. I address the unique needs of equivalence testing in the RD framework by providing an asymptotic test for the continuity of the regression function of pretreatment covariates at the cutoff. I also develop a test of continuity in the density of the forcing variable, using a scale-free asymptotic test that addresses the need for a way to define an equivalence range for the density test that can be easily interpreted by researchers.

A known drawback of equivalence tests is that they are very sensitive to the researcher specified equivalence range, unlike the standard null hypothesis test, a major advantage of which is a universally agreed upon null. To address this concern, I recommend use of an equivalence confidence interval, which is invariant to the equivalence range defined by the researcher, rather than reliance on *p*-values. The size of the equivalence confidence interval, rather than the statistical significance of the hypothesis test, can be used by researchers and readers to evaluate the strength of the design—a narrow range indicates limited evidence that the data are inconsistent with a valid design. In order to address bandwidth selection, the most consequential decision for any RD analysis, I rely on local linear estimators with optimal bandwidth selection (Calonico, Cattaneo, and Titiunik 2014; De la Cuesta and Imai 2016; Cattaneo, Jansson, and Ma 2019). These data-driven methods minimize discretion on the part of the researcher in determining an appropriate bandwidth. Cattaneo *et al*. (2020) note that "all predetermined covariates and placebo outcomes should be analyzed in the same way as the outcome of interest." They also discuss additional tests for sensitivity to bandwidth selection that apply to the statistical tests described in this manuscript.

The importance of statistical tests as evidence of a credible design is evident in the recent debate over the validity of the RD design for evaluating party incumbency advantage. Originally discussed in Lee, Moretti, and Butler (2004) and Lee (2008) in the context of US House Elections, recent scholars have called in to question the validity of the RD design for close elections (Snyder 2005; Caughey and Sekhon 2011; Grimmer *et al*. 2011) based on theoretical and statistical arguments. Follow-up studies by Eggers *et al*. (2015b) and De la Cuesta and Imai (2016) evaluate both US House elections and elections across the globe using novel data and analysis techniques, concluding that the statistical evidence generally favors close elections as a valid RD design. I return to this debate and reanalyze the statistical evidence in favor of the close elections RD design, ultimately finding more mixed evidence, particularly for certain geographies, than the recent studies would suggest.

2 Notation

Following the notation in Calonico *et al*. (2014), define the triple $(Y_i(1), Y_i(0), X_i)$ as the potential outcome under treatment, potential outcome under control, and the "forcing" variable, respectively, for individual i, where we observe a random sample of units $i = 1, 2, \ldots, n$. Assume that X_i has density $f(x)$. Define a treatment variable T_i, for which $T_i = 1(X_i \geq c)$ where $1(\cdot)$ is an indicator function for whether a unit is treated.[3] Since individual i can only be assigned to one treatment

3 This is commonly referred to as a sharp RD design. In a fuzzy RD, there is a discontinuity in $\Pr(T_i = 1 \mid X_i)$ at the cutoff, but treatment assignment is not a deterministic function of the forcing variable. I focus on the sharp RD in this paper, but similar logic can be applied to the falsification tests for fuzzy RD.

condition, we only observe one of the potential outcomes, specifically the potential outcome for control for all units below the cutoff and the potential outcome for treatment for units above. Define the observed outcome $Y_i = Y_i(1) * T_i + Y_i(0) * (1 - T_i)$ where we observe the random vector (Y_i, X_i) for each unit.

The identifying assumption necessary for RD design is continuity in the conditional expectation function of the potential outcomes at the cutoff, that is, $\mathbb{E}[Y_i(0) \mid X_i]$ and $\mathbb{E}[Y_i(1) \mid X_i]$ are continuous at the cutoff $X_i = c$. Formally, define the regression function of Y_i as $\mu_Y(x) = \mathbb{E}[Y_i \mid X_i = x]$, with $\mu_{Y+} = \lim_{x \to c+} \mu_Y(x)$ defined as the limit of $\mu_Y(x)$ from the right side of the cutoff, and $\mu_{Y-} = \lim_{x \to c-} \mu_Y(x)$ as the limit of $\mu_Y(x)$ from the left side of the cutoff, where μ_{Y+} and μ_{Y-} are both observable. Under continuity, the average treatment effect at the cutoff $\tau = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = c] = \mu_{Y+} - \mu_{Y-}$ is identified as the difference in the limits of the regression functions of the potential outcomes at the cutoff c (Hahn *et al*. 2001).

While continuity in the regression function for the potential outcomes is minimally sufficient for identification of τ at the cutoff, it is inherently untestable because we can never simultaneously observe both the potential outcomes above and below the cutoff. Lee (2008) employs a slightly different set of assumptions in which the density of the forcing variable is continuous at the cutoff for every individual i.[4] The assumption rules out "sorting" of individuals around the cutoff. Sorting is a concern if individuals can exert influence over which side of the cutoff they fall, and thus their treatment status. If sorting exists, we would observe a discontinuity in the density of the forcing variable. This sorting can be based on observable or unobservable characteristics, and leads to selection bias. Sorting can also lead to a discontinuity in the regression function for the potential outcomes, the minimal identifying assumption, if individuals sort on factors related to those potential outcomes, and therefore sorting can bias the estimation of τ.

The advantage of the Lee (2008) formulation is that the identifying assumption has to two testable implications in the data: (1) that the limits for the regression function of *any* pretreatment covariate Z_i, $\mu_Z = \mathbb{E}[Z_i \mid X_i]$, is continuous at the cutoff and (2) that the density of the forcing variable X_i, $f(x)$, is continuous at the cutoff. It is these two observable implications that are the focus of this paper. For the test of continuity in a pretreatment covariate, I will focus on the estimated difference in the regression function estimates for a pretreatment covariate Z on either side of the cutoff. Define $\mu_{Z+} = \lim_{x \to c+} \mu_Z(x)$ and $\mu_{Z-} = \lim_{x \to c-} \mu_Z(x)$, with $\tau_Z = \mu_{Z+} - \mu_{Z-}$. The method will test if τ_Z is sufficiently close to zero. For continuity in the density of the forcing variable I will consider the relationship of $f^- = \lim_{x \to c-} f(x)$ and $f^+ = \lim_{x \to c+} f(x)$, and test if the ratio of these two quantities is sufficiently close to 1. Details of these tests are provided in Sections 3.2 and 3.3, respectively.

3 Equivalence Testing in the RDD Setting

I now turn to the general form that a falsification test for a causal design should take, followed by a discussion of how to structure falsification tests for RD designs in the equivalence framework. There are key decisions a researcher must make when setting up an equivalence test, such as the definition of a "substantively inconsequential" equivalence range. I pay particular attention to the impact of this decision and how researchers should define this range. Importantly, I discuss the equivalence confidence interval, which is invariant to the researcher specified equivalence range, and serves as a transparent measure by which researchers and readers can evaluate evidence of equivalence.

4 Formally, define W_i to be a characteristic of unit i. Define (X_i, W_i) drawn jointly, with X_i observed and W_i unobserved, and define $F(x \mid w)$, the cdf of X conditional on W, such that $0 < F(x = c \mid W_i) < 1$ and $F(x \mid w)$ is continuously differentiable in X at the cutoff for any w in the support of W, with $f(x = c) > 0$. In this formulation, both the potential outcomes and the forcing variable can be a function of W, thus making W a potential confounder.

A statistical test should be structured such that the data has an opportunity to reject the null. The implication of Equation (1), for RD designs is that we should conduct falsification tests by starting with a null that there exists a discontinuity in the regression function for a pretreatment covariate, or the density of the forcing variable, at the cutoff and only reject this null if the data allow. This can be accomplished using equivalence tests, a type of statistical test in which the null hypothesis is that a parameter is of interest is proximally distant from zero against the alternative that it is within a small range around zero. Equivalence tests have a long history in biostatistics (Berger and Hsu 1996; Wellek 2010), and have been extended to balance and placebo tests used to justify "as-if" random causal designs (Rosenbaum *et al*. 2010; Hartman and Hidalgo 2018).[5]

While Equation (1) provides a theoretical structure for falsification tests for RD designs, we must construct a statistical test as a statement about a population parameter. For example, to test continuity in a pretreatment variable, Z, we want to test if the limits of the regression function are similar, or "equivalent," from above (μ_{Z+}) and below (μ_{Z-}) the cutoff. A researcher would set up a hypothesis test such that:

$$\begin{aligned} H_0 &: \tau_Z \geq \epsilon_U \quad or \quad \tau_Z \leq \epsilon_L \\ H_1 &: \epsilon_L < \tau_Z < \epsilon_U, \end{aligned} \tag{2}$$

where $[\epsilon_L, \epsilon_U]$ encodes the range in which τ_Z is sufficiently small, or the estimates of μ_{Z+} and μ_{Z-} are "substantively equivalent." When the hypothesis test is structured this way, a skeptical researcher will only reject a null of an important difference in favor of substantively inconsequential difference with sufficient evidence from the data. This is consistent with the goals of hypothesis testing, and prevents researchers from the temptation to incorrectly conflate lack of statistical power with a lack of substantive difference.

For example, in the close elections literature, it is common to test for continuity in the lagged vote margin. A researcher might argue that an equivalence range of ±2.5 percentage points is unlikely to significantly bias an observed effect of +5 percentage points in the election at time $t+1$. The equivalence test would then be structured with an equivalence range of $[-2.5, 2.5]$. If she observed a difference of +1 percentage point, the test in Equation (2) asks "can the data reject the null that the true difference in lagged vote margin is outside ±2.5 percentage points?" In practice, given the sensitivity of the test to the specified equivalence range, I argue researchers should instead focus on the size of the equivalence confidence interval, described in the next section, rather than directly specify the equivalence range. Details for the tests above are described in Section 3.2, where we will return to this example. An analogous test for continuity of the density of the forcing variable is described in Section 3.3.

3.1 The Equivalence Confidence Interval and the Equivalence Range

The most important decision a researcher must make when conducting an equivalence test is defining the equivalence range $[\epsilon_L, \epsilon_U]$. This is the range within which differences between two population parameters are considered substantively inconsequential, or equivalent. The p-value associated with the equivalence test is very sensitive to the researcher defined equivalence range. If a researcher specifies a large range, then she, all else equal, is more likely to reject the null in favor of equivalence. Conversely, a very small range is less likely to result in rejection of the null of a consequential difference. Small p-values resulting from large equivalence ranges carry less information about the credibility of the design than small p-values from a test with a narrow equivalence range. Given the considerable researcher degree

5 Equivalence testing is also appropriate when studying outcomes, particularly when studying "negligible," or substantively insignificant, effects (Wellek 2010; Gross 2014; Rainey 2014).

of freedom in defining the equivalence range, disagreement among researchers about an appropriate range is likely to arise.

To address the importance of defining the equivalence range, I argue researchers should instead focus on the equivalence confidence interval. This interval is akin to a confidence interval in the traditional test of difference—it determines the smallest equivalence range, supported by the data, that would reject a null of difference at the prespecified α-level.[6] Importantly, this interval is invariant to the prespecified equivalence range. Therefore, the equivalence confidence interval provides a transparent measure that researchers can substantively defend as inconsequential for possible bias. An advantage of the equivalence confidence interval is that, once the researcher has chosen the α-level at which she wishes to control the type I error, she does not need to directly specify the equivalence range in order to evaluate the hypothesis. The confidence interval can also help researchers avoid some common misunderstandings and misinterpretations of hypothesis tests and *p*-values (Gill 1999).

If the researcher does desire a decision rule for when she can reject a null of difference, then she must specify an equivalence range. When using equivalence tests for the purpose of falsification testing, the equivalence range should be constructed such that differences within the range are unlikely to cause substantial bias. By forcing researchers to consider, ex-ante, what differences are inconsequential enough so as allay concerns about bias, researchers must take time to carefully consider, and defend, what they consider substantively inconsequential. While implications of the identifying assumption are that $\tau_Z = 0$ and $f^- = f^+$, it is impossible to prove continuity with finite data, so researchers should pick a sufficiently small range that parsimoniously considers bias and power. While it is not possible to bound potential bias without additional assumptions, theory and substantive knowledge should be used to convincingly argue what levels of observed difference are tolerable. For example, a researcher could argue that a trivial discontinuity of one quarter of a percentage point in previous vote share is unlikely to indicate significant bias when estimating the impact of party incumbency, especially if estimated effects are sizeable.

An alternative approach for defining the equivalence range is the sensitivity approach. For example, similar to the assumption often used in sensitivity analyses, if there is a perfect linear relationship between imbalance in the election outcome at $t - 1$ and bias at $t + 1$, then an equivalence range can be defined based on the size of the observed effect. If the researcher observes an effect at $t + 1$ of 5pp, any difference in the equivalence range of ±2.5pp would be insufficient to reduce the effect to zero. Hartman and Hidalgo (2018) provide an in depth discussion of alternative approaches to defining the equivalence range, including the sensitivity approach as well as default values suggested in the literature.

As an example of the advantage of the equivalence confidence interval, when considering the placebo test for lagged vote share using the US House of Representatives data from 1880 to 2010, the point estimate is 0.16 percentage points, indicating that the incumbent party was 0.16 percentage points more likely to have won the election in time $t - 1$. Using an equivalence range of ±2.5 percentage points, the test can reject the null of a substantively large difference in favor of equivalence with a *p*-value of 0.04. The equivalence confidence interval is ±1.35 percentage points, indicating the data would reject the null with an equivalence range as small as ±1.35 percentage points at the $\alpha = 0.05$ level. Rather than argue that the data supports the claim of a valid design using a *p*-value of 0.04, which is sensitive to the choice of equivalence range, the researcher should argue that observable differences of less than ±1.35 percentage points are likely inconsequential for bias when estimating party incumbency advantage. The researcher should convincingly argue that the size of the equivalence confidence interval is negligible, rather than focusing on the *p*-value of the associated test.

6 A similar procedure is described in Section 19.3 of Rosenbaum *et al*. (2010).

3.2 A Statistical Test for the Continuity of the Regression Function for Pretreatment Covariates

Arguably the most important falsification test for RD designs is evidence of continuity of the regression function for highly predictive pretreatment covariates such as the lagged outcome. The test for continuity in the regression function of a pretreatment covariate uses the hypothesis test outlined in Equation (2). Recall that the equivalence range, $[\epsilon_L, \epsilon_U]$, is the range within which a researcher believes imbalances are substantively inconsequential. Obviously, any value of $\tau_Z > 0$ could be indicative of the failure of the identifying assumption, but finite samples require that we test a small range for the null hypothesis. The smaller the interval that the researcher specifies, the stronger her claim. Since the identifying assumption implies a difference of zero it is logical that the hypotheses are structured with symmetric bounds such that $\epsilon = \epsilon_U = -\epsilon_L$ for $\epsilon > 0$.[7]

Wellek (2010) outlines the general form for an asymptotic test for hypotheses of this type, which require a consistent estimator for τ_Z and $SE(\tau_Z)$. In this equivalence t-test, the null hypothesis in Equation (2) rejected using the following decision rule:

$$\text{Reject } H_0 \; iff \; |\hat{\tau}_Z / \widehat{SE}(\tau_Z)| < C_\alpha(\epsilon / \widehat{SE}(\tau_Z)), \tag{3}$$

where $C_\alpha(\psi)$ is the square root of the α-quantile of a χ^2-distribution with a single degree of freedom and noncentrality parameter ψ^2 (Wellek 2010).[8] One argument the researcher must specify is α in order to construct an equivalence confidence interval or conduct the hypothesis test. As when conducting any hypothesis test, the researcher should choose α to control the type I error at an acceptable level. I focus on the conventional, if arbitrary, level of α=0.05 in this manuscript, but the recent literature has proposed alternative values given the common misinterpretation of p-values (e.g., Benjamin *et al*. 2018).

To estimate $\hat{\tau}_Z$ and $\widehat{SE}(\tau_Z)$, I suggest the consistent estimators provided by Calonico *et al*. (2014).[9] As suggested by the RD literature, these estimators use optimal, data-driven methods for bandwidth selection. Bandwidth selection is one of the most important decisions for RD analyses, and these mean-squared error optimal estimators provide a principled, objective way of selecting the bandwidth. The same criterion for determining the bandwidth, such as mean-squared error optimal, should be used for falsification testing and outcome estimation, however the resulting bandwidth need not be the same (Cattaneo and Vazquez-Bare 2016).

Returning to the lagged vote margin example from Section 3, the researcher observes a difference of +1 percentage point with an estimated standard error of 0.5. The equivalence confidence interval is estimated as ±1.82 percentage points, and the researcher can transparently defend this interval as substantively inconsequential. If, instead, a researcher defines an equivalence range of ±2.5 percentage points, she can test the null hypothesis, at the $\alpha = 0.05$ level, that the lagged vote margin is outside this range. She finds that $1/0.5 = 2 < 3.35 = C_{0.05}(2.5/0.5)$, so she can reject the null of a large substantive difference in favor of the alternative of equivalence. The advantage of the equivalence confidence interval is that she can transparently defend this interval as substantively inconsequential rather than defending the choice of ±2.5 percentage points for the equivalence range.

Asymptotically, the equivalence t-test converges on the interval inclusion test, in which the researcher tests if a $100(1-2\alpha)$% confidence interval is entirely contained within the equivalence

7 Theoretical arguments about how direction of imbalance drives bias may allow a researcher to define a nonsymmetrical range.

8 Another way to construct this equivalence test is to used standardized units for ϵ, rather than on the raw scale of the covariate, with critical values adjusted appropriately.

9 Note that the Calonico *et al*. (2014) estimators impose additional estimation assumptions on the higher moments of the outcome and finite variance. $\hat{\tau}_Z$ is estimated using the bias-corrected estimators in Calonico *et al*. (2014), and $\widehat{SE}(\tau_Z)$ is estimated using the authors' robust variance estimator. Theorem 1 of the original manuscript shows these estimators meet the conditions of Wellek (2010).

range. Berger and Hsu (1996) show that the interval inclusion method corresponds to the Two-One-Sided Test (TOST) approach for equivalence, in which the two component null hypotheses outlined in Equation (2) are tested individually using one-sided *t*-tests, and the null is rejected if each of the composite nulls is rejected. While the interval inclusion method is the uniformly most powerful test (Romano 2005), the test outlined in Equation (3) is more powerful in finite samples than the interval inclusion method; the two tests converge quickly as sample size increases. Local polynomial estimators put very little weight on observations far from the cutoff, so the effective sample sizes near the cutoff are sometimes very small in RD estimation, making the additional power for the test described in Equation (3) particularly important in the RD context.

For simplicity, I focus on continuity in a single pretreatment outcome. However, the more variables for which a researcher can show continuity in the regression function, and the more predictive these variables are of the potential outcomes, the more credible the argument that the design is not affected by unobservable confounders.[10] Eggers *et al*. (2015b) and De la Cuesta and Imai (2016) discuss the use of multiple testing corrections when conducting multiple balance tests, and the *p*-values from equivalence tests can be directly input into these procedures.[11]

3.3 A Statistical Test for the Continuity in the Density of the Forcing Variable

As discussed in Section 2, under the assumptions described in Lee (2008), most researchers test for evidence of sorting by looking for a discontinuity in the density estimates of the forcing variable to the right and left of the cutoff. While the forcing variable can be related to the potential outcomes, the concern is that if units can exert control over which side of the cutoff they fall, then the potential outcomes may be discontinuous at the cutpoint. For example, as discussed by Caughey and Sekhon (2011), in competitive post-WWII house races, the incumbent party's candidate tends to have more political experience and resources. While this may not be enough to win an election, the concern is that, when the race is close, these individuals have a greater ability to exert control over the outcome by making maximal use of their resources, and they note that three-quarters of close races are won by the incumbent party. If these increased resources allow candidates to increase their probability of winning, and thus receiving treatment, it is possible that there is also a discontinuous jump in the future probability of a party winning. A valid design, which does not suffer from sorting, will exhibit continuity in the density of the forcing variable at the cutoff because units cannot exert control over the value of treatment.

McCrary (2008) was the first to propose a formal test for continuity in the density of the forcing variable. The McCrary test is conducted by first creating an under-smoothed histogram of points to the left and right of the cutoff, followed by local-linear estimation to smooth the histogram, once again conducted separately to the left and the right of the cutoff. The McCrary test, and subsequent related tests, require the researcher to set many nuisance parameters. To address this concern, Cattaneo *et al*. (2019) (CJM) have developed a nonparametric density estimator with boundary adaptive properties that can be used for consistently estimating the density at the cutoff. This method requires selection of only one parameter, the bandwidth. As with the continuity test, the CJM estimators select the bandwidth using a data-driven mean-squared error optimal method. Both the McCrary and CJM tests structure the null hypothesis such that the density is equal at the cutoff and use a traditional test-of-difference.

10 There is no strict guidance on how many variables researchers should test. The strength of the argument depends on the covariance structure of the variables and the potential outcomes. Even if strongly predictive variables show only slight discontinuities, the resulting bias could be of arbitrary magnitude or size.

11 Under current practice, there are problematic incentives for researchers to test numerous covariates, which would inflate multiple-testing *p*-values, and leads to ad-hoc problems such as "[A]dding irrelevant, and hence noisy, covariates can increase the number of hypotheses tested, thus reducing the probability of rejecting the null hypotheses." (De la Cuesta and Imai 2016, p. 389). Equivalence tests control the correct type I error, and therefore multiple testing methods can be applied without these concerns.

To formalize an equivalence test based on the density of the forcing variable, first note that equality of the density $f(x)$ at the cutoff implies that the ratio of the estimates should be equal (i.e., $f^+ = f^- \iff f^+/f^- = 1$). Unlike in the continuity test, where equivalence bounds can be defined on the scale of the pretreatment variable of interest (or in standardized units), constructing equivalence bounds on the scale of the probability density function is difficult, and highly dependent on the observed data. A formulation based on the ratio does not require knowledge of the range of the probability density function and still tests if there is a discontinuous jump in the relative likelihood of observing observations just above or just below the cutoff point.

The equivalence test for continuity in density is structured as:

$$\begin{aligned} H_0 &: \frac{f^+}{f^-} > \epsilon \quad \text{or} \quad \frac{f^+}{f^-} < \frac{1}{\epsilon} \\ H_A &: \frac{1}{\epsilon} < \frac{f^+}{f^-} < \epsilon. \end{aligned} \tag{4}$$

When testing a ratio, the equivalence range is symmetric on the ratio scale.[12] Rearranging each of the statements in the null as a one-sided test against zero, we can test the composite null using the intersection-union principle (Berger and Hsu 1996). The statistical test is then structured such that:

$$\text{Reject } H_0 \quad iff \quad T_1 \geq z_\alpha \quad \text{and} \quad T_2 \leq -z_\alpha, \tag{5}$$

where

$$T_1 = \frac{f^+ - \frac{1}{\epsilon} f^-}{\sqrt{\text{Var}(f^+) + \frac{1}{\epsilon^2}\text{Var}(f^-)}} \text{ and } T_2 = \frac{f^+ - \epsilon f^-}{\sqrt{\text{Var}(f^+) + \epsilon^2 \text{Var}(f^-)}}$$

and z_α is a standard normal critical value. Cattaneo *et al*. (2019) provide consistent estimators for $\hat{f^*}$ and $\widehat{SE}(f^*)$ that use mean-squared error optimal bandwidth selection.

To eliminate the need to prespecify the equivalence range, particularly given the lack of existing guidance on how to specify an appropriate equivalence range for the density function, I suggest the researcher focus on the equivalence confidence interval when providing evidence of credibility of the design, rather than the *p*-value associated with a prespecified equivalence range. The researcher should convincingly argue that the resulting range is narrow enough to assuage concerns about sorting.

Evidence against sorting strengthens the argument that there are no unmeasured confounders that could invalidate the design. Even if the data does not allow for rejection of the null in the equivalence density test, the researcher may still be able to argue the design is credible, particularly if there is strong evidence of continuity for covariates that are highly predictive of the potential outcomes. Ultimately, it is incumbent on the researcher to convincingly argue why remaining confounding is unlikely.

4 Simulations

Before turning to an application, I present simulations of the tests described in Sections 3.2 and 3.3. The simulations are intended to capture both discontinuous jumps in the regression function for a pretreatment covariate and discontinuous jumps in the density of the forcing variable around

12 Note that if we observe a 20% decrease from $f^+ = 0.1$ to $f^- = 0.08$, then if we define our test statistic as f^+/f^- we get $0.1/0.08 = 1.25$, however if we define the test statistic as f^-/f^+, then $0.08/0.1 = 0.8$. Another way to consider this is that the log of the test would give us a symmetric test of the difference in the density estimates.

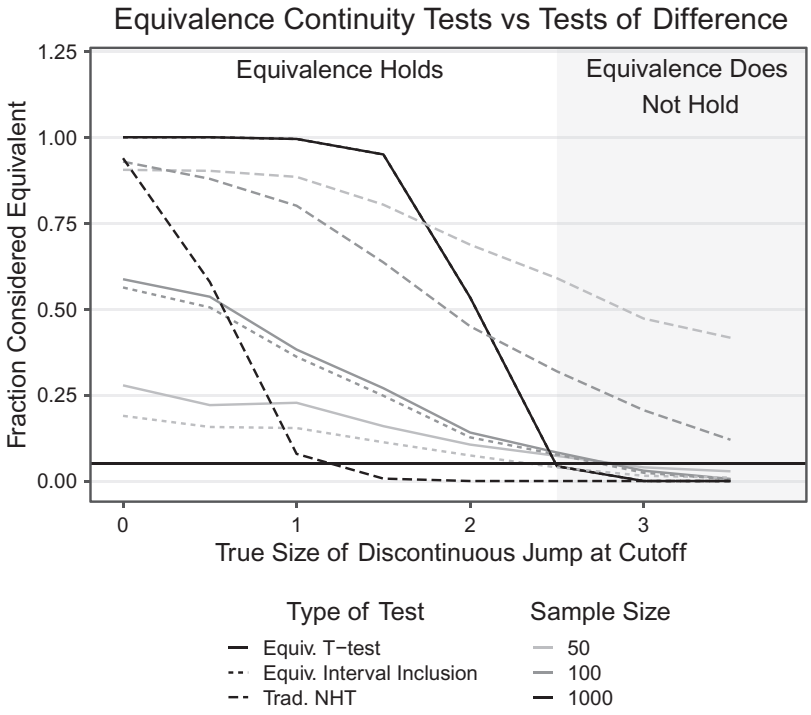

Figure 1. Equivalence Tests vs. Tests of Difference for tests in continuity. Solid lines present the rejection rate for the equivalence test, which shows maximal power when continuity holds, and maintains at least nominal level when the true difference is outside the prespecified equivalence range.

the cutoff. The data-generating process is similar to the simulations found in Calonico *et al*. (2014), which mimics the Lee (2008) data. I induce jumps in the density following the simulations in Cattaneo *et al*. (2019). The simulations resemble close elections, where the forcing variable is two-party vote share (in percentage points), and in this case Z, the pretreatment covariate, is two-party vote share in the previous election. I conduct 1,000 simulations for each parameter value described below. Details about the data generating process can be found in the supplementary materials.[13]

I first consider the test for continuity in the regression function for Z. Results are presented for discontinuous jumps in Z of $\tau_Z \in \{0, 1.25, 2.5, 3.75\}$ (presented on the x-axis in Figure 1, with an equivalence range of ± 2.5).[14] In the context of close elections the simulations are modeled after, for example, this would correspond to an equivalence range of ± 2.5 percentage points in the previous election vote share. Using the equivalence t-test for continuity described in Equation (3), I plot the proportion of simulations that reject at the $\alpha = 0.05$ level and compare the performance with two methods—the interval inclusion method for equivalence, and the traditional null hypothesis test-of-difference. The interval inclusion and traditional test-of-difference are conducted using the estimators discussed in Calonico *et al*. (2014). While it is a misinterpretation of the test, consistent with common practice when using tests-of-difference as a falsification test, I consider a p-value of greater than 0.05 as evidence of equivalence.[15] There is no sorting in any of these simulations.

There are three important results in Figure 1. First, the equivalence t-test for continuity (solid line) performs as expected, reaching power of near one when continuity holds, and rejecting no more than 100(α)% of the time, with $\alpha = 0.05$ represented by the horizontal line, when equivalence does not hold (the gray region). Second, the interval inclusion method (short dash)

13 Replication files for this manuscript are available at https://doi.org/10.7910/DVN/IVRHIR (Hartman 2020).
14 The range of Z is about 41–53, which is similar to previous vote share seen in close election RD designs.
15 Results are substantively similar if researchers adjust for their criterion for equivalence and use a p-value of 0.15, such as suggested by Cattaneo, Frandsen, and Titiunik (2015) on p. 9.

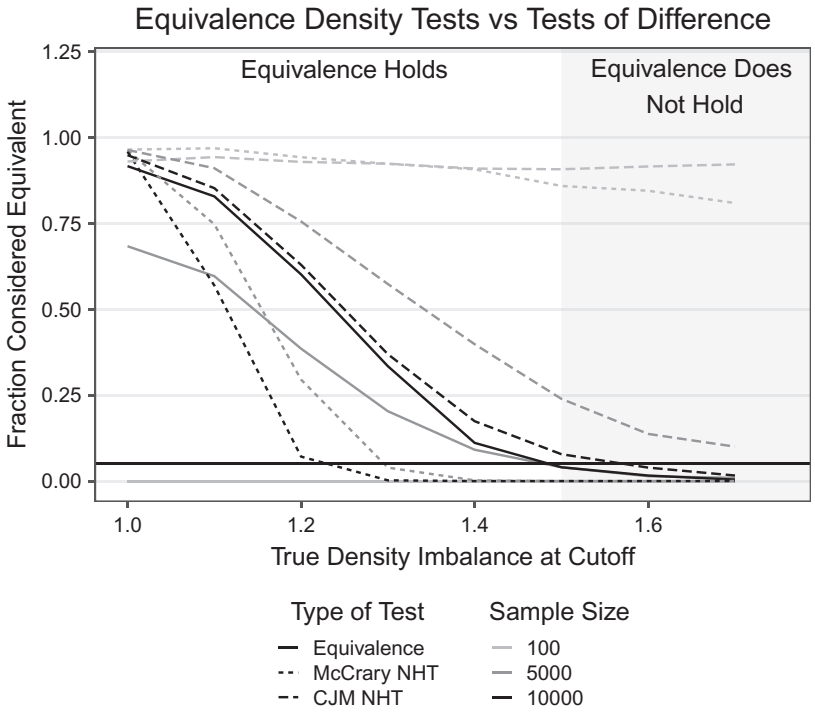

Figure 2. Equivalence tests versus tests of difference for tests of continuity in density. Solid lines present the equivalence test which shows maximal power when continuity holds, and maintains at least nominal level when the true imbalance is outside the prespecified equivalence range.

performs similarly to the equivalence *t*-test with larger sample sizes, but is less powerful in smaller sample sizes.

Finally, the poor performance of a test-for-difference (long dash) is clear. Power is limited when the design is valid and continuity exists (i.e., there is no discontinuous jump) because the traditional null hypothesis test will still reject 5% of the time. This problem will be exacerbated if researchers consider larger *p*-value cutoffs. With small sample sizes, the test frequently erroneously indicates that there is evidence of equivalence (i.e., large *p*-values). In the gray shaded region, where equivalence does not hold, there is a high fraction considered equivalent by the traditional null hypothesis test when the sample size is small, but not by the equivalence tests.

Next I turn to the test for evidence of sorting. I show the performance of the equivalence density test (solid line) described in Equation (5) using an equivalence range of $[2/3, 1.5]$[16]; results show the proportion of simulations that reject at the $\alpha = 0.05$ level. I compare this to two tests of difference—the McCrary density test (McCrary 2008) (small dash) and the density test discussed in Cattaneo *et al*. (2019) (referred to as the CJM test). For both of these tests, I consider a *p*-value of greater than 0.05 as evidence of equivalence. No discontinuous jump in the regression function of Z is considered in these simulations.

Figure 2 presents results from the simulations for continuity in the density of the forcing variable. Along the *x*-axis is the true density ratio at the cutoff, which ranges from 1, indicating no sorting, to 1.75, indicating a jump of 75% in the probability of receiving treatment right at the cutoff. True jumps in density of between $[2/3, 1.5]$ are considered "equivalent," as defined by the equivalence range, although I only plot values of one or larger.

16 There is no existing guidance on an appropriate range for a density function. The literature on bioequivalence for effectiveness of generic versus name-brand drugs uses a fairly strict range of [4/5, 5/4] (Berger and Hsu 1996). Given the power concerns with density estimation, and I use a wider interval of [2/3, 3/2]. Further practical guidance on how influential a discontinuity in the density function can is for bias in the RD setting, such as through sensitivity analyses for this parameter, is left as an area of future research.

Two important things are evident in Figure 2. First, the equivalence density test performs as expected, with power approaching 1 when the density is continuous, but maintaining nominal levels when the true jump is larger than the equivalence range. However, the density test has much lower power than the tests for continuity. This problem is not unique to the equivalence approach, as the difference-based approaches are also under-powered, as evidenced by the fact that the McCrary and CJM tests rarely detect sorting—whether or not equivalence holds—when sample sizes are small. This lack of power exhibited by the difference-based approaches should be particularly concerning given the widespread use tests-of-difference for evidence of sorting. Future work should focus on designing a more sensitive density test.

5 Application: Close Elections

First discussed by Lee (2008), one of the most widely used applications of the RD design is in the study of close elections. The design requires that, in a close election, candidates cannot precisely control whether they barely win or barely lose. The validity of this assumption came under question with the findings of Snyder (2005), Caughey and Sekhon (2011), and Grimmer *et al*. (2011), who provide theoretical and statistical evidence of candidate sorting, especially in post-WWII congressional races. Recently, De la Cuesta and Imai (2016) and Eggers *et al*. (2015b) find evidence suggesting that the close-election RD design appears, generally, credible.

While theoretical arguments against possible mechanisms for sorting are an important part for arguing the close election RD design is credible, most of the statistical evidence relies on traditional null hypothesis tests. To exhibit how to use equivalence-based falsification tests for RD designs, I reanalyze the global close elections data of Eggers *et al*. (2015a), focusing on lagged vote-share as the pretreatment outcome for testing for continuity.[17] Ultimately, I find that the statistical evidence in favor of the close-elections design is much more mixed than the recent literature would suggest. This evidence should be combined with theoretical arguments to carefully consider the validity of close-elections in any given geographic context.

I begin by replicating the falsification tests conducted in Eggers *et al*. (2015b), focusing on lagged vote share at $t-1$ as the primary covariate of interest for testing for continuity. The authors' argument that the RD design does not seem sensitive to manipulation in most electoral settings, despite findings to the contrary in the US House of Representatives, hinges on a series of statistically insignificant tests-of-difference for lagged vote share at time $t-1$.[18] I begin by conducting equivalence-based continuity tests on the lagged vote share for the winning party with an equivalence range of ± 2.5 percentage points. In the original analysis, the authors present a series of p-values using difference-in-means, local linear, and polynomial regressions in various windows around the cutoff. They indicate that "results [are] not shown if there are insufficient data points within a given bandwidth, to avoid biased or uninformative inferences," acknowledging that small sample sizes may lead to conflation of insignificant difference with evidence of a valid design.

Figure 3 presents the results of the equivalence continuity tests.[19] The first thing to note is that there is no need to eliminate tests that are under-powered—the equivalence test will convey the additional uncertainty with wider equivalence confidence intervals, as is evident in the French

17 I also reanalyze the US Congressional data of Caughey and Sekhon (2011) in the supplementary materials using an equivalence-based continuity test, and show how to appropriately apply a multiple testing correction, as recommended by De la Cuesta and Imai (2016).

18 Table 4 in the original manuscript.

19 No equivalence confidence interval is presented for the US Mayors dataset. There are some scenarios in which the equivalence confidence interval is undefined, in particular when the test rejects at the α-level for a noncentrality parameter of zero (corresponding to an empty equivalence range). This occurs when $|t| < \sqrt{(\chi^2)^{-1}(\alpha, 1)}$, which implies that the test would reject for any level of ϵ, which defines the noncentrality parameter. For example, `qchisq(0.05, 1)` $= 0.0039 \implies |t| = 0.06$, or that the standard error is 15.9 times larger than the point estimate. In the US Mayors rate, the t stat is $-0.35/6.78 = -0.053$, driven primarily by a very large standard error. In these cases, which may occur with a small point estimate and large standard error estimate, I do not report an equivalence confidence interval. If the researcher wanted to report a conservative interval, they could report the equivalence confidence interval from the two-one-sided test.

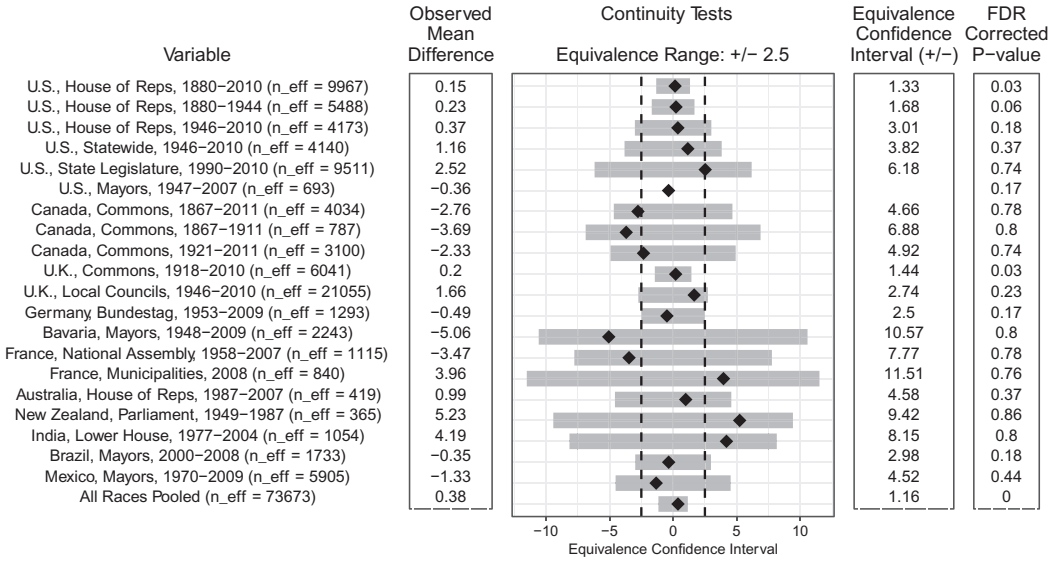

Figure 3. Equivalence test for continuity in the lagged vote share for the winning party as applied to the Eggers *et al*. (2015b) data. The vertical dashed line corresponds to the tested equivalence range of ±2.5 percentage points. Black diamonds correspond to the point estimate. Gray bars indicate the equivalence confidence interval. The *p*-value includes a false discovery rate correction. Continuity would imply the point estimate should be near 0.

Municipal, Australian House, and New Zealand Parliament elections, which were not presented in the original manuscript. Second, some analyses exhibit large point-estimates for difference, indicating the importance of focusing not only on the *p*-value of a test, but the estimate. For example, the India lower house races, which have a point estimate of 4.19, had *p*-values ranging from 0.2 to 0.89 in the original analysis, but fails to reject using the equivalence-based test.

Finally, results indicate that the evidence of the continuity of the regression function of the lagged outcome is mixed. Rather than focusing on *p*-values, which require agreement that an equivalence range of ±2.5 percentage points is substantively inconsequential, researchers should focus on the width of the equivalence confidence interval, which conveys the minimal equivalence range supported by the data. Some races have very small equivalence confidence intervals, such as the early US House of Representatives (1880–1944) which can support a minimum range of ±1.68 percentage points. However many races also have very wide equivalence confidence intervals, such as the French Municipal elections, which supports a minimum range of ±11.51 percentage points. Some of these wide intervals may be due to small sample size, but depending on which context a researcher is working in, they should carefully consider the evidence at hand. When combined, the overall equivalence confidence interval for all pooled races is ±1.16 percentage points, with a very small point-estimate of 0.38, which supports the Eggers *et al*. (2015b) argument that the RD design is "broadly applicable."

Next I reanalyze the tests for sorting around the cutoff across geographies using the equivalence-based test.[20] Results, presented in Figure 4, show that many, but not all, of the elections show little evidence of sorting, as defined by a 50% jump in the density estimate. I present the equivalence confidence interval for each race, showing the smallest range supported by the data at the $\alpha = 0.05$ level. The pooled election evidence indicates that close-election RD designs do not exhibit strong evidence of sorting, however some individual geographies cannot reject a null consistent with sorting indicating causal effect estimates in those contexts may not be credible. Using an equivalence range of $[2/3, 3/2]$, there is stronger evidence of a valid RD design across geographies than the continuity results using a ±2.5 percentage point range. In the

20 Table 5 of the original manuscript.

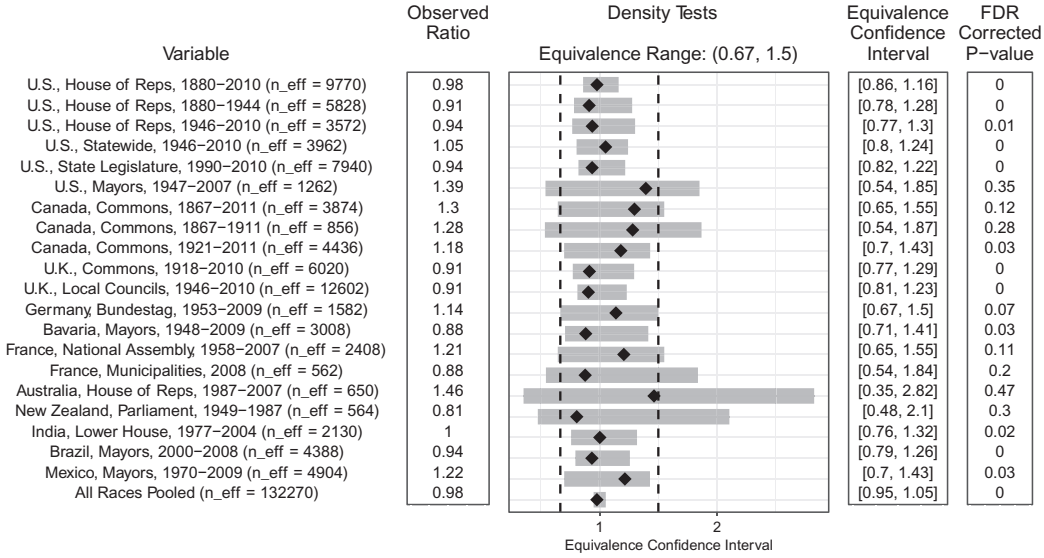

Figure 4. Equivalence test for no sorting as applied to the Eggers *et al*. (2015b) data. The vertical dashed line corresponds to the tested equivalence range of [2/3, 1.5]. Black diamonds correspond to the point estimate. Gray bars indicate the equivalence confidence interval, which are nonsymmetric in this case to account for the asymmetry of a ratio. The *p*-value includes a false discovery rate correction. No sorting would imply that the ratio of the density estimates should be near 1.

original analysis, all geographies for which results were presented had statistically insignificant results. In the reanalysis, while most geographies show evidence of no-sorting, the evidence is not consistent across all geographies.

Combining the results of the equivalence based continuity tests for pretreatment covariates and the density of the forcing variable provides a more mixed set of findings regarding statistical evidence in favor of the validity of individual electoral contexts for the RD design than in Eggers *et al*. (2015b) and De la Cuesta and Imai (2016). While the literature lays out compelling structural theories for why sorting is unlikely in electoral settings, statistical evidence from the equivalence tests, in which we aim to reject a null hypothesis that the data are *inconsistent* with a valid RD design, has mixed evidence supporting the design across geographies. While they do not lessen the need for strong theory of the mechanisms by which sorting can occur, the methods discussed here can help researchers provide statistical evidence that their RD design is credible. The results here suggest that a researcher should carefully evaluate the evidence of a valid RD for her specific geographic and temporal context. This can be done by providing evidence of numerous pretreatment covariates and testing for sorting. I provide an example in the supplementary materials of a specific context by reevaluating the results in Caughey and Sekhon (2011), in which I find statistical evidence supporting the authors' concern about RD analysis of the post-WWII US Congress.

6 Discussion and Conclusion

Falsification tests are an important tool for researchers when arguing for a valid, identified causal design. When a researcher cannot control the assignment mechanism in her study, causal identification will always rely on a set causal identification assumptions that cannot be tested directly with observed data. Theoretical arguments are important for justifying the validity of a design, but a researcher should also leverage the data, where possible, to bolster her claims about the credibility of the design. One way she can do this is by constructing statistical tests that only rejects the null the data is inconsistent with a flawed design when the data allow.

I have presented two such statistical tests here that are tailored for RD designs. Using the equivalence testing framework, I construct a test for continuity of the regression function of a

pretreatment covariate as well as a test for continuity in the density of the forcing variable at the cutoff. These tests require the researcher to specify, ex-ante, an equivalence range within which observed differences are substantively inconsequential, and the results provide evidence against a flawed design, as defined by this equivalence range. These tests are highly sensitive to the range that the researcher specifies, so I have also discussed the importance of the equivalence confidence interval, the minimum equivalence range supported by the data at the α-level, which serves as a transparent statistic a researcher should argue is sufficiently small to alleviate concerns about bias. Simulations compare the performance of these tests to the current practice tests-of-difference, showing the value of equivalence based tests in both small and large samples. I have focused on sharp RD designs in this manuscript, but similar tests are easily extended to fuzzy and kink designs. Publicly available code implementing these tests is available.[21]

These falsification tests will better allow researchers achieve their goals: rejecting the null in these tests provides statistical evidence that the data are consistent with a valid RD design. There is no concern about conflating lack of power with equivalence, and increasing the sample size will increase the power of the test without any ad-hoc justification of the interpretation of the *p*-value. This increases the likelihood that researchers pursue designs when they are valid, and that they put in the shoe-leather when the evidence is more mixed. As discussed in De la Cuesta and Imai (2016), multiple testing considerations are important. With equivalence-based tests, researchers can combine their analyses with standard multiple testing corrections or combined tests without increasing the likelihood of falsely considering a design valid, a concern with current practice.

When applied to the recent debate on the validity of the close election RDs, equivalence tests provide a murkier set of findings than recent studies. On average, when pooling close-elections across the globe, there is strong evidence of the validity of the close election RD design. However, in some specific geographic areas and time periods, where data are more limited, there may not be enough data to rule out sorting in a specific electoral context. For example, a reanalysis of the postwar data in US Congressional races finds evidence more consistent with the results of Caughey and Sekhon (2011), where many pretreatment covariates appear imbalanced, including incumbency at $t-1$. While these tests do not directly invalidate the RD designs in these contexts, it places increased burden on the researcher to rule out sorting using strong theory and alternative evidence in any given context. As discussed in Eggers *et al*. (2015b, p. 272), "extraordinary care is required in order to generate inferences given the presence of imbalances" but the data need not be dismissed completely in the face of failed falsification tests.

In this manuscript, I have focused on the traditional RD design framework. Recent literature has clarified the conditions under which the RD can be treated like a local experiment, where treatment is "as-if" randomly assigned in a narrow window around the cutoff. The main necessary identifying assumptions are continuity in the potential outcomes, and that the potential outcomes are constant (and therefore the regression function is flat) within a narrow bandwidth around the cutoff (De la Cuesta and Imai 2016; Sekhon and Titiunik 2017). The equivalence tests described in Hartman and Hidalgo (2018) directly apply to the difference-in-means estimator employed in falsification tests for this design.

A special consideration in the local randomization framework, though, is for bandwidth selection, an important decision for any RD design. Current best practice for window selection is discussed in Cattaneo *et al*. (2015), where they describe a procedure that finds the widest window where scores are unrelated to covariates within the window, but are associated outside of it; this is done using an exact test with a null of equality. They acknowledge that "our ultimate goal is to learn whether the data support the existence of a neighborhood around the cutoff where our null hypothesis fails to be rejected. In this sense, the roles of Type I and Type II error are interchanged

21 https://github.com/ekhartman/rdd_equivalence

in our context" (p. 9) and that this could be addressed using an equivalence test. Given the lack of power in small windows, equivalence-based falsification tests are most appropriate for window selection. Because the identifying assumption for the local experiment framework requires an association between the covariate and the forcing variable outside the window, researchers may wish to combine an equivalence test within and a test-of-difference outside to optimize selection of the window.

Falsification tests are an essential part of the toolkit for justifying causal identifying assumptions. Here I provide two such tests for RD designs that allow researchers to provide statistical evidence that their designs are consistent with empirically testable implications of the identifying assumptions. While the role of theory is not minimized when arguing for the validity of the design, these tests allow the data to strengthen the credibility of causal findings.

Data Availability Statement

The replication materials for this paper can be found at Hartman (2020).

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2020.43.

Acknowledgments

I would like to thank Graeme Blair, Naoki Egami, Danny Hidalgo, Jeff Lewis, and Santiago Olivella for their detailed feedback, as well as the participants of the Latin American PolMeth 2018 conference and PolMeth 2018.

References

Benjamin, D. J. et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2(1):6–10.

Berger, R., and J. Hsu. 1996. "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets." *Statistical Science* 11(4):283–302.

Calonico, S., M. D. Cattaneo, and R. Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regressiondiscontinuity Designs." *Econometrica* 82(6):2295–2326.

Cattaneo, M. D., B. R. Frandsen, and R. Titiunik. 2015. "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the US Senate." *Journal of Causal Inference* 3(1):1–24.

Cattaneo, M. D., N. Idrobo, and R. Titiunik. 2020. *A Practical Introduction to Regression Discontinuity Designs: Foundations. Elements in Quantitative and Computational Methods for the Social Sciences*. Cambridge: Cambridge University Press.

Cattaneo, M. D., M. Jansson, and X. Ma. 2019. "Simple local Polynomial Density Estimators." *Journal of the American Statistical Association* 115(531):1449–1455.

Cattaneo, M. D., and G. Vazquez-Bare. 2016. "The Choice of Neighborhood in Regression Discontinuity Design." *Observational Studies* 2(December):134–146.

Caughey, D., and J. S. Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close US House Races, 1942–2008." *Political Analysis* 19(4):385–408.

De la Cuesta, B., and K. Imai. 2016. "Misunderstandings about the Regression Discontinuity Design in the Study of Close Elections." *Annual Review of Political Science* 19:375–396.

Eggers, A. C., A. Fowler, J. Hainmueller, A. B. Hall, and J. M. Snyder. 2015a. "Replication Data for: On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races." *American Journal of Political Science* 59(1), 259–274.

Eggers, A. C., A. Fowler, J. Hainmueller, A. B. Hall, and J. M. Snyder. 2015b. "On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races." *American Journal of Political Science* 59(1):259–274.

Gill, J. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3):647–674.

Grimmer, J., E. Hersh, B. Feinstein, and D. Carpenter. 2011. "*Are Close Elections Random*?" Unpublished manuscript.

Gross, J. H. 2014. "Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance." *American Journal of Political Science* 59(3):775–788.

Hahn, J., P. Todd, and W. van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69(1):201–209.
Hartman, E. 2020. "Replication Data for: Equivalence Testing for Regression Discontinuity Designs." https://doi.org/10.7910/DVN/IVRHIR, Harvard Dataverse, V1.
Hartman, E., and F. D. Hidalgo. 2018. "An Equivalence Approach to Balance and Placebo Tests." *American Journal of Political Science* 62(4): 1000–1013.
Lee, D. S. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics* 142(2):675–697.
Lee, D. S., E. Moretti, and M. J. Butler. 2004. "Do voters Affect or Elect Policies? Evidence from the US House." *The Quarterly Journal of Economics* 119(3):807–859.
Lee, D. S. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics* 142(2):675–697.
McCrary, J. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142(2):698–714.
Rainey, C. 2014. "Arguing for a Negligible Effect." *American Journal of Political Science* 58(4):1083–1091.
Romano, J. P. 2005. "Optimal Testing of Equivalence Hypotheses." *The Annals of Statistics* 33(3):1036–1047.
Rosenbaum, P. R. et al. 2010. *Design of Observational Studies*, vol. 10. New York: Springer.
Sekhon, J. S., and R. Titiunik. 2017. "On Interpreting the Regression Discontinuity Design as a Local Experiment." *Advances in Econometrics* 38:1–28.
Skovron, C., and R. Titiunik. 2015. "A Practical Guide to Regression Discontinuity Designs in Political Science." *American Journal of Political Science* 2015:1–36.
Snyder, J. 2005. "Detecting Manipulation in US House Elections." Unpublished manuscript.
Thistlethwaite, D. L., and D. T. Campbell. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." *Journal of Educational Psychology* 51(6):309.
Wellek, S. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. 2nd edn. Boca Raton, FL: CRC Press.