TESTING IDENTIFYING ASSUMPTIONS IN FUZZY REGRESSION DISCONTINUITY DESIGNS*

YOICHI ARAI[a] YU-CHIN HSU[b] TORU KITAGAWA[c] ISMAEL MOURIFIÉ [d] YUANYUAN WAN[e]

ABSTRACT. We propose a new specification test for assessing the validity of fuzzy regression discontinuity designs (FRD-validity). We derive a new set of testable implications, characterized by a set of inequality restrictions on the joint distribution of observed outcomes and treatment status at the cut-off. We show that this new characterization exploits all of the information in the data that is useful for detecting violations of FRD-validity. Our approach differs from and complements existing approaches that test continuity of the distributions of running variables and baseline covariates at the cut-off in that we focus on the distribution of the observed outcome and treatment status. We show that the proposed test has appealing statistical properties. It controls size in a large sample setting uniformly over a large class of data generating processes, is consistent against all fixed alternatives, and has non-trivial power against some local alternatives. We apply our test to evaluate the validity of two FRD designs. The test does not reject FRD-validity in the class size design studied by Angrist and Lavy (1999) but rejects it in the insurance subsidy design for poor households in Colombia studied by Miller, Pinto, and Vera-Hernández (2013) for some outcome variables. Existing density continuity tests suggest the opposite in each of the two cases.

Keywords: Fuzzy regression discontinuity design, nonparametric test, inequality restriction, multiplier bootstrap.

Date: Saturday 20th March, 2021.

$*$. This paper merges and replaces the unpublished independent works of Arai and Kitagawa (2016) and Hsu, Mourifié, and Wan (2016).

a. School of Social Sciences, Waseda University, yarai@waseda.jp.

b. Institute of Economics, Academia Sinica; Department of Finance, National Central University; Department of Economics, National Chengchi University; CRETA, National Taiwan University, ychsu@econ.sinica.edu.tw.

c. Department of Economics, University College London, t.kitagawa@ucl.ac.uk.

d. Department of Economics, University of Toronto, ismael.mourifie@utoronto.ca.

e. Department of Economics, University of Toronto, yuanyuan.wan@utoronto.ca.

Acknowledgement: We thank Josh Angrist, Matias Cattaneo, Hide Ichimura, Robert McMillan, Phil Oreopoulos, Zhuan Pei, Christoph Rothe, Yuya Sasaki, and Marcos Vera-Hernandez for beneficial comments. We also thank the participants of the 2016 Asian, the 2017 North American and the European Meetings of the Econometric Society, the Online Causal Inference Seminar, the Cemmap Workshop on Regression Discontinuity and on Advances in Econometrics, the Workshop on Advances in Econometrics at Dogo, Tsinghua Workshop on Econometrics, and the department seminars at Duke University, PSU, UBC, the Queen's University, University of Arizona, and Tokyo University of Science. We are grateful to Jeff Rowley for excellent research assistance. Financial support from the ESRC through the ESRC Centre for Microdata Methods and Practice (CeMMAP) (grant number RES-589-28-0001), the ERC through the ERC starting grant (grant number 715940), the Japan Society for the Promotion of Science through the Grants-in-Aid for Scientific Research No. 15H03334, Ministry of Science and Technology of Taiwan (MOST107-2410-H-001-034-MY3), Academia Sinica Taiwan (CDA-104-H01 and :AS-IA-110-H01), Center for Research in Econometric Theory and Applications (107L9002) from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education of Taiwan, and Waseda University Grant for Special Research Projects is gratefully acknowledged.

1. Introduction

Regression discontinuity (RD) design, first introduced by Thistlethwaite and Campbell (1960), is one of the most widely used quasi-experimental methods in program evaluation studies. The RD design exploits discontinuity in treatment assignment due to administrative or legislative rules based on a known cut-off of an underlying assignment variable, which we refer to as the running variable. The RD design is called sharp if the probability of being treated jumps from zero to one, and is called fuzzy otherwise. See Imbens and Lemieux (2008), and Lee and Lemieux (2010) for reviews, and Cattaneo and Escanciano (2017) for recent advances of the literature.

The RD design identifies the causal impact of the treatment by comparing the outcomes of treated and non-treated individuals close to the cut-off. The validity of the RD design relies crucially on the assumption that those individuals immediately below the cut-off have the same distribution of unobservables as those individuals immediately above the cut-off. The first formalization of this argument appears in Hahn, Todd, and Van der Klaauw (2001, HTV hereafter), which utilises a potential outcomes framework to establish identification of causal effects at the cut-off. Subsequently, Frandsen, Frölich, and Melly (2012, FFM hereafter), Dong and Lewbel (2015), and Cattaneo, Keele, Titiunik, and Vazquez-Bare (2016) consider a refined set of identifying conditions. In the fuzzy regression discontinuity design (FRD) setting, the two key conditions for identification, which we refer to as *FRD-validity*, are (i) *local continuity*, the continuity of the distributions of the potential outcomes and treatment selection heterogeneity at the cut-off, and (ii) *local monotonicity*, the monotonicity of the treatment selection response to the running variable at the cut-off.

The credibility of FRD-validity is controversial in many empirical contexts. For instance, agents (or administrative staff) may manipulate the value of the running variable to be eligible for their preferred treatment. If their manipulation depends on their underlying potential outcomes, this can lead to a violation of the local continuity condition. Even when manipulation of the running variable is infeasible or absent, the local continuity condition can fail if the distribution of unobservables is discontinuous across the cut-off. This is a common concern in empirical research, for instance, when the RD design exploits geographical boundaries across which individuals are less likely to relocate but the ethnic distribution of the population is changing discontinuously. See Dell (2010), and Eugster, Lalive, Steinhauer, and Zweimüller (2017). As a related example, violation of local continuity becomes a concern when multiple programs share an index of treatment assignment and its threshold (e.g. the poverty line, state borders, etc.), but an individual's treatment status is observed only for the treatment of interest. See Miller, Pinto, and Vera-Hernández (2013), Carneiro and Ginja (2014), and Keele and Titiunik (2015) for examples and discussions of this issue.

Motivated by a clearer economic interpretation and the availability of testable implications, Lee (2008) imposes a *stronger* set of identifying assumptions that implies continuity of the distributions of the running variable and covariates at the cut-off. Following his approach, researchers routinely assess the continuity condition by applying the tests of McCrary (2008), Otsu, Xu, and Matsushita (2013), Cattaneo, Jansson, and Ma (2020), and Canay and Kamat (2018). When the running variable is manipulated, Gerard, Rokkanen, and Rothe (2020) provides a partial identification approach in the presence of "one-sided manipulation." As noted by McCrary (2008), however, in the absence of Lee's additional identifying assumption, the continuity of the distributions of the running variable and baseline covariates at the cut-off is neither necessary nor sufficient for FRD-validity, and rejection or acceptance of the existing tests is not informative about FRD-validity or violation thereof.

This paper proposes a novel test for FRD-validity. We first derive a new set of testable implications, characterized by a set of inequality restrictions on the joint distribution of observed outcomes and treatment status at the cut-off. These testable implications are necessary conditions for FRD-validity, while we show they are sharp in the sense that they cannot be strengthened without additional assumptions. We propose a nonparametric test for these testable implications. The test controls size uniformly over a large class of distributions of observables, is consistent against all fixed alternatives violating the testable implications, and has non-trivial power against some local alternatives. Implementability and asymptotic validity of our test neither restricts the support of Y nor presumes continuity of the running variable's density at the cut-off.

The testable implication that our test assesses differs from and complements the testable implication that the existing density continuity tests focus on. As we illustrate in Section 2.3, there are important empirical contexts where the results of the existing approach are not informative about FRD-validity while ours are. They include scenarios where the distribution of unobservables is discontinuous at the cut-off, multiple programs share the same running variable and the same threshold, and manipulation of the running variable is driven by factors independent of the potential outcomes. It is also important to note that our testable implication assesses local monotonicity, about which the existing approach is not informative. The novelty of our approach is that it exploits those aspects of the data that are informative in assessing FRD-validity but that have been neglected by the existing density continuity approach. We therefore recommend that our test is implemented alongside existing tests for continuity of the running variable density, regardless of the results thereof.

To illustrate our proposal, we apply our test to the designs studied in Angrist and Lavy (1999), and Miller, Pinto, and Vera-Hernández (2013). Angrist and Lavy (1999) use the discontinuity of class size with respect to enrollment due to Maimonides' rule to identify the causal effect of class size on student performance. We do not find statistically significant violation of our new testable implication for FRD-validity for any of the four outcome variables (Grade 4 Math and Verb, Grade 5 Math and Verb). In contrast, the existing continuity test suggests statistically significant evidence for discontinuity of the running variable's density at the cut-off (see Otsu, Xu, and Matsushita, 2013). Miller, Pinto, and Vera-Hernández (2013) evaluate the impact of "Colombia's Régimen Subsidiado"– a publicly financed insurance program–on 33 outcomes, where program eligibility is determined by a poverty index. Since our approach makes use of observations of not only the running variable but also of treatment status and the observed outcome, it has the unique feature of being outcome-specific, i.e. when multiple outcomes are studied within the same FRD design, researchers can assess credibility of FRD-validity separately for each outcome variable. In this example, the continuity test supports continuity of the running variable density at the cut-off, while we find statistically significant evidence for the violation of our new testable implication for FRD-validity for 3 outcome variables (Household Education Spending, Total Spending on Food, and Total Monthly Spending). This result suggests further investigation would be beneficial for identifying and estimating the causal effect on these outcomes.

The rest of the paper is organized as follows. In Section 2, we lay out the main identifying assumptions that our test aims to assess and derive their testable implications. Section 3 provides test statistics and shows how to obtain their critical values. Monte Carlo experiments shown in Section 4 examine the finite sample performance of our tests. Section 5 presents the empirical applications. Section 6 concludes the paper. The Supplemental Material (Arai, Hsu, Kitagawa, Mourifié, and Wan (2020)) provides detailed discussion of how our test differs and complements existing tests, several extensions, the asymptotic validity of our test, all proofs, and additional empirical results.

2. Identifying Assumptions and Sharp Testable Implications

2.1. Setup and Notation.

We adopt the potential outcome framework introduced in Rubin (1974). Let $(\Omega, \mathcal{F}, P)$ be a probability space, where we interpret Ω as the population of interest and $\omega \in \Omega$ as a generic individual in the population.

Let R be an observed continuous random variable with support $\mathcal{R} \subset \mathbb{R}$.[1] We call R the *running variable*. Let $D(\cdot,\cdot) : \mathcal{R} \times \Omega \to \{0,1\}$ and $D(r,\omega)$ be the *potential treatment* that individual ω would have received, had her running variable been set to r. For $d \in \{0,1\}$, we define mappings $Y_d(\cdot,\cdot) : \mathcal{R} \times \Omega \to \mathcal{Y} \subset \mathbb{R}$ and let $Y_d(r,\omega)$ denote the *potential outcome* of individual ω had her treatment and running variable been set to d and r, respectively.

We view $(Y_1(r,\cdot), Y_0(r,\cdot), D(r,\cdot))_{r\in\mathcal{R}}$ as random elements indexed by r and write them as $(Y_1(r), Y_0(r), D(r))$ when it causes no confusion. By definition, $D(R) \in \{0,1\}$ is the observed treatment and we abbreviate it as D. Likewise, we denote the observed outcome by $Y = Y_1(R)D(R) + Y_0(R)(1 - D(R))$ throughout the paper. We use P to denote the joint distribution of $\big((Y_1(r), Y_0(r), D(r))_{r\in\mathcal{R}}, R\big)$, which induces the joint distribution of observables (Y, D, R).[2] We assume throughout that the conditional distribution of (Y, D) given $R = r$ is well-defined for all r in some neighborhood of r_0, and that $\lim_{r\downarrow r_0} D(r)$ and $\lim_{r\uparrow r_0} D(r)$ are well defined for all ω. Note that by letting the potential outcomes be indexed by r, we allow the running variable to have a direct causal effect on outcomes. This could be relevant in some empirical applications as discussed in Dong and Lewbel (2015), and Dong (2018).

Analogous to the local average treatment effect (LATE) framework (Imbens and Angrist (1994)), we define the compliance status $T(r,\omega)$ of individual ω in a small neighborhood of the cut-off r_0 based on how the potential treatment varies with r. Similar to FFM, Bertanha and Imbens (2020), and Dong and Lewbel (2015), for $\epsilon > 0$, we classify the population members into one of the following

[1] In this paper we consider a continuous running variable. Kolesár and Rothe (2018) study inference on ATE in the sharp regression discontinuity designs with a discrete running variable.

[2] For the purpose of exposition, we do not introduce other observable covariates X here. Appendix C.2 of the Supplemental Material incorporates X into the analysis.

five categories:

$$
T_\epsilon(\omega) = \begin{cases} \mathbf{A}, & \text{if } D(r,\omega) = 1, \text{ for } r \in (r_0 - \epsilon, r_0 + \epsilon), \\ \mathbf{C}, & \text{if } D(r,\omega) = 1\{r \geq r_0\}, \text{ for } r \in (r_0 - \epsilon, r_0 + \epsilon), \\ \mathbf{N}, & \text{if } D(r,\omega) = 0, \text{ for all } r \in (r_0 - \epsilon, r_0 + \epsilon), \\ \mathbf{DF}, & \text{if } D(r,\omega) = 1\{r < r_0\}, \text{ for } r \in (r_0 - \epsilon, r_0 + \epsilon), \\ \mathbf{I}, & \text{otherwise} \end{cases}, \tag{1}
$$

where **A, C, N, DF** and **I** represent "always takers", "compliers", "never takers", "defiers" and "indefinite", respectively.[3]

2.2. **Identifying Assumptions and Testable Implication.** We present the main identifying assumptions and their testable implications. In the statement of the assumptions we assume that all the limiting objects exist.

Assumption 1 (Local monotonicity)**.** *For* $t \in \{\mathbf{DF}, \mathbf{I}\}$, $\lim_{\epsilon \to 0} P(T_\epsilon = t | R = r_0 + \epsilon) = 0$ *and* $\lim_{\epsilon \to 0} P(T_\epsilon = t | R = r_0 - \epsilon) = 0$.

Assumption 2 (Local continuity)**.** *For* $d = 0, 1$, $t \in \{\mathbf{A}, \mathbf{C}, \mathbf{N}\}$, *and any measurable subset* $B \subseteq \mathcal{Y}$,

$$
\lim_{\epsilon \to 0} P(Y_d(r_0 + \epsilon) \in B, T_\epsilon = t | R = r_0 + \epsilon) = \lim_{\epsilon \to 0} P(Y_d(r_0 - \epsilon) \in B, T_\epsilon = t | R = r_0 - \epsilon).
$$

Assumptions 1 and 2 play similar roles to the instrument monotonicity and instrument exogeneity (exclusion and random assignment) assumptions in the LATE framework. Assumption 1 says that as the neighborhood of r_0 shrinks, the conditional proportion of defiers and indefinites converges to zero, implying that only "always takers", "compliers", and "never takers" may exist at the limit. The local continuity assumption says that the conditional joint distributions of potential outcomes and compliance types are continuous at the cut-off. Our local continuity condition concerns distributional continuity rather than only continuity of the conditional mean (and so is unlike HTV).

The main feature of FRD designs is that the probability of receiving treatment is discontinuous at the cut-off. To be consistent with the local monotonicity assumption, we specify the discontinuity so that the propensity score jumps *up* as r goes above the cut-off.

[3]The above definition coincides with the definition of types in FFM as $\epsilon \to 0$. As pointed out by Dong and Lewbel (2015), for a given ϵ and a given individual ω, this definition implicitly assumes the group to which ω belongs does not vary with r. This way of defining the treatment selection heterogeneity does not restrict the shape of $P(D = 1 | R = r)$ over $(r_0 - \epsilon, r_0 + \epsilon)$.

Assumption 3 (Discontinuity)**.** $\pi^{+} \equiv \lim_{r \downarrow r_0} P(D=1|R=r) > \lim_{r \uparrow r_0} P(D=1|R=r) \equiv \pi^{-}$.

Under Assumptions 1 to 3, the compliers' potential outcome distributions at the cut-off, defined as

$$\begin{aligned} F_{Y_1(r_0)|\mathbf{C},R=r_0}(y) &\equiv \lim_{r \to r_0} P\left(Y_1(r) \leq y | T_{|r-r_0|} = \mathbf{C}, R=r\right), \\ F_{Y_0(r_0)|\mathbf{C},R=r_0}(y) &\equiv \lim_{r \to r_0} P\left(Y_0(r) \leq y | T_{|r-r_0|} = \mathbf{C}, R=r\right), \end{aligned}$$

are identified by the following quantities:[4] for all $y \in \mathcal{Y}$,

$$\begin{aligned} F_{Y_1(r_0)|\mathbf{C},R=r_0}(y) &= \frac{\lim_{r \downarrow r_0} \mathbb{E}_P[1\{Y \leq y\}D|R=r] - \lim_{r \uparrow r_0} \mathbb{E}_P[1\{Y \leq y\}D|R=r]}{\pi^{+} - \pi^{-}}, \\ F_{Y_0(r_0)|\mathbf{C},R=r_0}(y) &= \frac{\lim_{r \uparrow r_0} \mathbb{E}_P[1\{Y \leq y\}(1-D)|R=r] - \lim_{r \downarrow r_0} \mathbb{E}_P[1\{Y \leq y\}(1-D)R=r]}{\pi^{+} - \pi^{-}}. \end{aligned}$$

This is analogous to the distributional identification result by Imbens and Rubin (1997) for the LATE model. The identification of the compliers' potential outcome distributions implies the identification of a wide class of causal parameters including the average effect amongst the compliers and local quantile treatment effects.[5] Our identification result modifies FFM's Lemma 1 to accommodate the fact that we do not exclude r from the potential outcomes.

Note that Assumption 3 can be tested using the inference methods proposed by Calonico, Cattaneo, and Titiunik (2014), and Canay and Kamat (2018). We therefore focus on testing Assumptions 1 and 2.

The next theorem shows that *local monotonicity* and *local continuity* together imply a set of inequality restrictions on the distribution of data.

Theorem 1. *(i) Under Assumptions 1 and 2, the following inequalities hold:*

$$\lim_{r \uparrow r_0} \mathbb{E}_P[1\{y \leq Y \leq y'\}D|R=r] - \lim_{r \downarrow r_0} \mathbb{E}_P[1\{y \leq Y \leq y'\}D|R=r] \leq 0 \tag{2}$$

$$\lim_{r \downarrow r_0} \mathbb{E}_P[1\{y \leq Y \leq y'\}(1-D)|R=r] - \lim_{r \uparrow r_0} \mathbb{E}_P[1\{y \leq Y \leq y'\}(1-D)|R=r] \leq 0 \tag{3}$$

[4]For completeness, we show this identification result in Proposition E.1 in Appendix E.2 of the Supplemental Material.

[5]Assumptions 1 and 2 play similar roles to FFM's Assumptions I3 and I2, respectively. The main difference from FFM's assumptions is that FFM define the compliance status solely at the limit, and assume that the conditional distributions of the potential outcomes given the limiting compliance status and the running variable are continuous at the cut-off.

for all $y, y' \in \mathbb{R}$.

(ii) For a given distribution of observables (Y, D, R), *assume that the conditional distribution of* Y *given* (D, R) *has a probability density function with respect to a dominating measure* μ *on* $\mathcal{Y}$, *has an integrable envelope with respect to* μ, *and whose left-limit and right-limit with respect to the conditioning variable* R *are well defined at* $R = r_0$, $\mu-a.s.$ *If inequalities (2) and (3) hold, there exists a joint distribution of* $(\tilde{D}(r), \tilde{Y}_1(r), \tilde{Y}_0(r) : r \in \mathcal{R})$ *such that Assumptions 1 and 2 hold, and the conditional distribution of* $\tilde{Y} = \tilde{Y}_1(R)\tilde{D}(R) + \tilde{Y}_0(R)(1 - \tilde{D}(R))$ *and* $\tilde{D} = \tilde{D}(R)$ *given* $R = r$ *induces the conditional distribution of* (Y, D) *given* $R = r$ *for all* $r \in \mathcal{R}$.

Theorem 1 (i) shows a necessary condition that the distribution of observable variables has to satisfy under the FRD-validity conditions. In other words, a violation of inequalities (2) and (3) is informative that at least one of the FRD-validity conditions is violated. Theorem 1 (ii) clarifies that inequalities (2) and (3) are the most informative way to detect all of the observable violations of the FRD-validity assumptions and the testable implications cannot be strengthened without making further assumptions. We emphasize, however, that FRD-validity is a refutable but not a *confirmable* assumption, i.e., finding inequalities (2) and (3) hold in data does not guarantee FRD-validity.

Similar to the testable implications of the LATE model considered in Balke and Pearl (1997), Imbens and Rubin (1997), Heckman and Vytlacil (2005), Kitagawa (2015), and Mourifié and Wan (2017), the testable implications of Theorem 1 (i) can be interpreted as an FRD version of *non-negativity of the potential outcome density functions for the compliers at the cut-off.* Despite such an analogy, the framework and features specific to RD designs give rise to some important differences and challenges. First, the assumption that we test is continuity of the conditional distributions of the potential outcomes and compliance status local to the cut-off, rather than the global exclusion or no-defier restrictions of the standard LATE model. Second, since the testable implications concern distributional inequalities local to the cut-off, the construction of the test statistic requires proper smoothing with respect to the conditioning running variable.

2.3. **How does our testable implication differ from existing implications?** FRD-validity, as defined by Assumptions 1 and 2, does not constrain the marginal density of R to be continuous at the cut-off. This contrasts with the testable implications of continuity of the running variable and covariate densities obtained in Lee (2008), and Dong (2018), which hinge on a stronger restriction such that the density of the running variable given the potential outcomes is continuous at the cut-off. See McCrary (2008), Otsu, Xu, and Matsushita (2013), Cattaneo, Jansson, and Ma (2020), and Bugni and Canay (2018) for tests of the continuity of the running variable density, and Canay and Kamat (2018) for tests of the continuity of the covariate densities.

The testable implication of Theorem 1 (i) is valid no matter whether one assumes such an additional restriction or not. The testable implication concerns the joint distribution of (Y, D) local to the cut-off, which the existing approach of assessing continuity of the densities of the running variable and observable covariates does not make use of. In this sense, our approach, which does not require continuity of the running variable's density, complements the existing approach of using continuity tests and we recommend the implementation of our test (proposed below) in any FRD studies, whatever results the existing continuity tests yield.

There are several important empirical contexts where supporting or rejecting continuity of the running variable's density is not informative about FRD-validity, while the testable implication of Theorem 1 (i) can be. First, even when the running variable's density is known to be continuous, it is still often controversial to assume that the distribution of unobservable heterogeneity affecting the outcomes is continuous at the cut-off. For instance, when an RD design exploits geographical or language boundaries (e.g., Dell (2010), and Eugster, Lalive, Steinhauer, and Zweimüller (2017)), the distribution of (unobservable) ethnicity may change discontinuously, even though individuals are distributed smoothly over the space. If the discontinuity of the distribution of unobservables leads to violation of the testable implication of Theorem 1 (i), our approach correctly refutes FRD-validity.

Second, if multiple programs share the same running variable and the same threshold (compound treatments), an FRD design that ignores the other programs can lead to violation of continuity of the potential outcome distributions (for the program of interest), even when the density of the running variable is continuous. For instance, empirical scenarios that rely on a spatial regression discontinuity design exploiting jurisdictional, electoral, or market boundaries (see Keele and Titiunik (2015), and references therein) can violate local continuity in this way. The issue of compound treatments is also of concern when multiple social programs targeted at the poor assign their eligibility according to a common poverty index and poverty line. (Carneiro and Ginja (2014)).

Third, in contrast to the previous two contexts, discontinuity of the running variable's density does not necessarily imply violation of local continuity if manipulation of the running variable is independent of the potential outcomes (possibly conditional on observable covariates). In this case, the testable implication of Theorem 1 (i) does not refute FRD-validity even though the running variable's density is discontinuous. See our empirical application in Section 5.1, below. In addition, Appendix B of the Supplemental Material provides detailed analytical comparisons between the testable implications of Lee (2008) and ours.

Another distinguishing feature of our approach is that our testable implication can also detect violation of local monotonicity. It is therefore valuable to assess the testable implication also in those scenarios where local continuity is credible while local monotonicity is less credible. Examples include studies examining the returns to field of study or college major, exploiting discontinuity generated by a centralized score-based admission system (Hastings, Neilson, and Zimmerman (2014), and Kirkeboen, Leuven, and Mogstad (2016)). In this context, the validity of local monotonicity can be a concern if an individual's choice of treatment (e.g., graduating with a degree in science rather than a degree in humanities) is different from their initial assignment in the program (e.g., admitted to a science or humanities program). Defiers can exist if some students always switch from their assigned major based on revisions of their beliefs or preferences.[6]

3. Testing Procedure

This section proposes a testing procedure for the testable implications of Theorem 1 (i). We assume that a sample consists of independent and identically distributed (i.i.d.) observations, $\{(Y_i, D_i, R_i)\}_{i=1}^n$. Noting that the inequality restrictions of Theorem 1 (i) amount to an infinite number of unconditional moment inequalities local to the cut-off, we adopt and extend the inference procedure for conditional moment inequalities developed in Andrews and Shi (2013) by incorporating the local feature of the RD design.[7] The implementation and asymptotic validity of our test neither restricts the support of Y nor presumes continuity of the running variable's density at the cut-off. See Appendix D of the Supplemental Material for regularity conditions and the asymptotic validity of our test.

[6] See Zafar (2011), and Stinebrickner and Stinebrickner (2014) for empirical evidence on how college students form and revise their beliefs on own academic outcomes for their majored and non-majored subjects and how this relates to their subsequent switch of majors.

[7] Other approaches and recent advances of the inference of conditional moment inequalities include Chernozhukov, Lee, and Rosen (2013), Armstrong and Chan (2016), and Chetverikov (2018). The methods proposed in these works are free from the infinitesimal uniformity factor η in Algorithm 1. Formal investigation of their applicability to the current regression discontinuity context is beyond the scope of this paper.

Consider a class of instrument functions $\mathcal{G}$ indexed by $\ell \in \mathcal{L}$:

$$
\begin{aligned}
&\mathcal{G} = \left\{g_\ell(\cdot) = 1\{\cdot \in C_\ell\} : \ell \equiv (y, y') \in \mathcal{L}\right\}, \text{ where} \\
&C_\ell = [y, y'] \cap \mathcal{Y}, \\
&\mathcal{L} = \{(y, y') : -\infty \le y \le y' \le \infty\}.
\end{aligned}
$$

$\mathcal{G}$ consists of indicator functions of closed and connected intervals on $\mathcal{Y}$. Expressing the inequalities (2) and (3) by

$$
\begin{aligned}
\nu_{P,1}(\ell) &\equiv \lim_{r\uparrow r_0} \mathbb{E}_P[g_\ell(Y)D|R=r] - \lim_{r\downarrow r_0} \mathbb{E}_P[g_\ell(Y)D|R=r] \le 0, \\
\nu_{P,0}(\ell) &\equiv \lim_{r\downarrow r_0} \mathbb{E}_P[g_\ell(Y)(1-D)|R=r] - \lim_{r\uparrow r_0} \mathbb{E}_P[g_\ell(Y)(1-D)|R=r] \le 0,
\end{aligned} \tag{4}
$$

for all $\ell \in \mathcal{L}$, we set up the null and alternative hypotheses as

$$
\begin{aligned}
&H_0 : \ \nu_{P,1}(\ell) \le 0 \text{ and } \nu_{P,0}(\ell) \le 0 \text{ for all } \ell \in \mathcal{L}, \\
&H_1 : \ H_0 \text{ does not hold.}
\end{aligned} \tag{5}
$$

Noting that H_0 is equivalent to $\sup_{d\in\{0,1\},\ell\in\mathcal{L}} \omega_d(\ell)\nu_{P,d}(\ell) \le 0$ for a positive weight function $\omega_d(\ell) > 0$, we construct our test statistic by plugging in estimators of $\nu_{P,d}(\ell)$ weighted by the inverse of its standard error estimate.

We construct $\hat{\nu}_d(\ell)$, an estimator for $\nu_{P,d}(\ell)$, as the difference of the two local linear regressions estimated from below and above the cut-off. We do not vary the bandwidths over $\ell \in \mathcal{L}$, but we allow them to vary across the cut-offs; let $h_+ = c_+h$ and $h_- = c_-h$ be the bandwidths above and below the cut-off, respectively. We assume that their convergence rates with respect to the sample size n are common, as specified by h, e.g., $h = n^{-1/4.5}$. The difference between h_+ and h_- can be captured by possibly distinct constants c_+ and c_-.

Let $\sigma_{P,d}(\ell)$ be the asymptotic standard deviation of $\sqrt{nh}(\hat{\nu}_d(\ell) - \nu_{P,d}(\ell))$ and $\hat{\sigma}_d(\ell)$ be a uniformly consistent estimator for $\sigma_{P,d}(\ell)$. See Algorithm 1, below, for its construction. To ensure uniform convergence of the variance weighted processes, we weigh $\hat{\nu}_d(\ell)$ by a trimmed version of the standard error estimators, $\hat{\sigma}_{d,\xi}(\ell) = \max\{\xi, \hat{\sigma}_d(\ell)\}$, where $\xi > 0$ is a trimming constant chosen by the user. See footnote 9 for the choice of ξ in our simulation study. We then define a

Kolmogorov-Smirnov (KS) type test statistic,

$$\widehat{S}_n = \sup_{d \in \{0,1\},\ \ell \in \mathcal{L}} \frac{\sqrt{nh} \cdot \hat{v}_d(\ell)}{\hat{\sigma}_{d,\xi}(\ell)}. \tag{6}$$

A large value of $\widehat{S}_n$ is statistical evidence against the null hypothesis. The cardinality of $\mathcal{L}$ is infinite if Y is continuously distributed, while with our construction of $\hat{v}_d(\ell)$ and $\hat{\sigma}_{d,\xi}(\ell)$ shown in Appendix A, we can coarsen $\mathcal{L}$ to the class of intervals spanned by the observed values of Y in the sample,

$$\hat{\mathcal{L}} \equiv \left\{ [Y_i, Y_j] : Y_i \le Y_j, i, j \in \{1, \ldots n\} \right\}, \tag{7}$$

without changing the value of the test statistic. In the Monte Carlo studies of Section 4 and the empirical applications of Section 5, we standardize and rescale the range of Y to the unit interval (by applying a transformation through the cdf of the standard normal distribution $\Phi(\cdot)$),[8] and employ the following coarsening of the class of intervals:

$$\mathcal{L}_{coarse} = \Big\{ (y, y+c) :\ c^{-1} = q, \text{ and } q \cdot y \in \{0, 1, 2, \cdots, (q-1)\} \text{ for } q = 1, 2, \cdots, Q \Big\}. \tag{8}$$

As done in Hansen (1996), and Barrett and Donald (2003) in different contexts, we obtain asymptotically valid critical values by approximating the null distribution of the statistic using multiplier bootstrap. Algorithm 1, below, summarizes the implementation of our test. Theorems D.1-D.3 in Appendix D of the Supplemental Material show that the proposed test controls size at pre-specified significant levels uniformly, rejects fixed alternatives with probability approaching one, and has good power against a class of local alternatives.

Algorithm 1. (Implementation)

i. Specify a finite class of intervals $\mathcal{L}^*$. For instance, $\mathcal{L}^* = \hat{\mathcal{L}}$ of (7), or a coarsened version with the standardized outcome, $\mathcal{L}^* = L_{coarse}$ of (8) with a choice of finite integer Q (e.g., $Q = 15$).

ii. For each $\ell \in \mathcal{L}^*$, let $\hat{m}_{1,+}(\ell)$ and $\hat{m}_{1,-}(\ell)$ be local linear estimators for $\lim_{r \downarrow r_0} \mathbb{E}_P[g_\ell(Y)D|R = r]$ and $\lim_{r \uparrow r_0} \mathbb{E}_P[g_\ell(Y)D|R = r]$, respectively. Similarly, let $\hat{m}_{0,+}(\ell)$ and $\hat{m}_{0,-}(\ell)$ be local linear estimators for $\lim_{r \downarrow r_0} \mathbb{E}_P[g_\ell(Y)(1-D)|R = r]$ and $\lim_{r \uparrow r_0} \mathbb{E}_P[g_\ell(Y)(1-D)|R = r]$,

[8] Since the null hypothesis and the test statistic are invariant to strictly monotonic transformations of Y, this standardization does not affect the theoretical guarantee and the empirical results of our test.

respectively. See equation (11) in Appendix A for their closed-form expressions. Obtain $\hat{\nu}_1(\ell)$ and $\hat{\nu}_0(\ell)$ as follows:

$$\hat{\nu}_1(\ell) = \hat{m}_{1,-}(\ell) - \hat{m}_{1,+}(\ell), \quad \hat{\nu}_0(\ell) = \hat{m}_{0,+}(\ell) - \hat{m}_{0,-}(\ell). \tag{9}$$

iii. For each $\ell \in \mathcal{L}^*$, calculate sample analogs of the influence functions

$$\hat{\phi}_{\nu_1,i}(\ell) = \sqrt{nh}\left(w_{n,i}^- \cdot (g_\ell(Y_i)D_i - \hat{m}_{1,-}(\ell)) - w_{n,i}^+ \cdot (g_\ell(Y_i)D_i - \hat{m}_{1,+}(\ell))\right),$$
$$\hat{\phi}_{\nu_0,i}(\ell) = \sqrt{nh}\left(w_{n,i}^+ \cdot (g_\ell(Y_i)(1-D_i) - \hat{m}_{0,+}(\ell)) - w_{n,i}^- \cdot (g_\ell(Y_i)(1-D_i) - \hat{m}_{0,-}(\ell))\right),$$

where the definitions of the weighting terms $\{(w_{n,i}^+, w_{n,i}^-) : i = 1, \ldots, n\}$ are given in Appendix A. We then estimate the asymptotic standard deviation $\sigma_{P,d}(\ell)$ by $\hat{\sigma}_d(\ell) = \sqrt{\sum_{i=1}^n \hat{\phi}^2_{\nu_d,i}(\ell)}$ and obtain the trimmed estimators as $\hat{\sigma}_{d,\xi}(\ell) = \max\{\xi, \hat{\sigma}_d(\ell)\}$.[9]

iv. Calculate the test statistic $\widehat{S}_n = \widehat{S}_n = \sup_{d\in\{0,1\},\ \ell\in\mathcal{L}^*} \frac{\sqrt{nh}\cdot\hat{\nu}_d(\ell)}{\hat{\sigma}_{d,\xi}(\ell)}$.

v. Let a_n and B_n be sequences of non-negative numbers. For $d = 0, 1$ and $\ell \in \mathcal{L}$, define $\psi_{n,d}(\ell)$ as

$$\psi_{n,d}(\ell) = -B_n \cdot 1\left\{\frac{\sqrt{nh}\cdot\hat{\nu}_d(\ell)}{\hat{\sigma}_{d,\xi}(\ell)} < -a_n\right\}. \tag{10}$$

Following Andrews and Shi (2013, 2014), we use $a_n = \sqrt{0.3\ln(n)}$ and $B_n = \sqrt{\frac{0.4\ln(n)}{\ln\ln(n)}}$.

vi. Draw $U_1, U_2, \cdots U_n$ as i.i.d. standard normal random variables that are independent of the original sample. Compute the bootstrapped processes, $\widehat{\Phi}_{\nu_1}(\ell)$ and $\widehat{\Phi}_{\nu_0}(\ell)$, defined as

$$\widehat{\Phi}_{\nu_1}(\ell) = \sum_{i=1}^n U_i \cdot \hat{\phi}_{\nu_1,i}(\ell), \quad \widehat{\Phi}_{\nu_0}(\ell) = \sum_{i=1}^n U_i \cdot \hat{\phi}_{\nu_0,i}(\ell).$$

vii. Iterate Step (vi) $\bar{B}$ times ($\bar{B}$ is a large integer) and denote the realizations of the bootstrapped processes by $\left(\widehat{\Phi}^b_{\nu_1}(\cdot), \widehat{\Phi}^b_{\nu_0}(\cdot) : b = 1, \ldots, \bar{B}\right)$. Let $\hat{q}(\tau)$ be the τ-th empirical quantile of $\left\{\sup_{d\in\{0,1\},\ \ell\in\mathcal{L}^*}\left\{\frac{\widehat{\Phi}^b_{\nu_d}(\ell)}{\hat{\sigma}_{d,\xi}(\ell)} + \psi_{n,d}(\ell)\right\} : b = 1, \ldots, \bar{B}\right\}$. For significance level $\alpha < 1/2$, obtain

[9] In the simulations, we set $\xi = \sqrt{a(1-a)}$, where $a = 0.0001$. We also use $a \in \{0.001, 0.03, 0.5\}$. The results are insensitive to the choice of a. These tuning parameters are motivated by the observation that the denominator of the asymptotic variance takes the form of $p_\ell(1-p_\ell)$, where $p_\ell = \lim_{r\to r_0}\mathbb{P}(Y \in C_\ell, D = d|R = r)$.

a critical value $\hat{c}_\eta(\alpha)$ of the test by $\hat{c}_\eta(\alpha) = \hat{q}(1-\alpha+\eta)+\eta$, where $\eta > 0$ is an arbitrarily small positive number, e.g., 10^{-6}.[10]

viii. Reject H_0 if $\widehat{S}_n > \hat{c}_\eta(\alpha)$.

Following the existing papers in the moment inequality literature, Step vii in Algorithm 1 uses the generalized moment selection (GMS) proposed by Andrews and Soares (2010), and Andrews and Shi (2013). It is similar to the recentering method of Hansen (2005), and Donald and Hsu (2016), and the contact set approach of Linton, Song, and Whang (2010).

Regarding the bandwidths for the local linear estimators in step ii, our informal recommendation is to have the bandwidth of $\hat{m}_{d,+}(\ell)$, $d = 1, 0$, common for all $\ell \in \mathcal{L}^*$ and the bandwidth of $\hat{m}_{d,-}(\ell)$, $d = 1, 0$, common for all $\ell \in \mathcal{L}^*$. We denote the two bandwidths by h_+ and h_-, respectively, and allow $h_+ \neq h_-$. There is merit to using the bandwidths that are recommended for point estimation of the LATE at the cut-off, such as the bandwidths suggested in Imbens and Kalyanaraman (2012), Calonico, Cattaneo, and Titiunik (2014), and Arai and Ichimura (2016). This is because the FRD-Wald estimator is numerically equal to the difference of the means between the following distribution function estimates for compliers:

$$\begin{aligned}
\hat{F}_{Y_1(r_0)|\mathbf{C},R=r_0}(y) &= \frac{\hat{m}_{1,+}((-\infty,y)) - \hat{m}_{1,-}((-\infty,y))}{\hat{\pi}^+ - \hat{\pi}^-}, \\
\hat{F}_{Y_0(r_0)|\mathbf{C},R=r_0}(y) &= \frac{\hat{m}_{0,-}((-\infty,y)) - \hat{m}_{0,+}((-\infty,y))}{\hat{\pi}^+ - \hat{\pi}^-},
\end{aligned}$$

where $\hat{m}_{1,+}((-\infty,y))$ and $\hat{m}_{0,+}((-\infty,y))$ use h_+, $\hat{m}_{1,-}((-\infty,y))$ and $\hat{m}_{0,-}((-\infty,y))$ use h_-, and $\hat{\pi}^+$ and $\hat{\pi}^-$ are the local linear estimators for $\lim_{r\downarrow r_0} P(D=1|R=r)$ and $\lim_{r\uparrow r_0} P(D=1|R=r)$ with bandwidths h_+ and h_-, respectively. Accordingly, reusing these bandwidths to compute our test statistic, we assess nonnegativity of the compliers' potential outcome densities based on the same in-sample information as that which the point estimate for the compliers' causal effect relies on.[11]

[10] This η constant is called an infinitesimal uniformity factor and is introduced by Andrews and Shi (2013) to avoid the problems that arise due to the presence of the infinite-dimensional nuisance parameters $\nu_{P,1}(\ell)$ and $\nu_{P,0}(\ell)$.

[11] Alternatively, we may want to choose bandwidths so as to optimize a power criterion. We leave power-optimizing choices of bandwidth for future research. Algorithm 1 provides some default choices for other tuning parameters, ξ, a_n, and B_n, without claiming that these choices are optimal. According to our Monte Carlo studies and empirical applications considered in Sections 4 and 5, the test results are not sensitive to mild departures from the default choices.

4. SIMULATION

This section investigates the finite sample performance of the proposed test by Monte Carlo experiments. We consider six data generating processes (DGPs) including two DGPs, Size1-Size2, for examining the size properties and four DGPs, Power1-Power4, for examining the power properties of the test. For all DGPs, we set the cut-off point at $r_0 = 0$.

4.1. Size properties.

Size1 Let $R \sim N(0,1)$ truncated at -2 and 2. The propensity score $P(D=1|R=r) = 0.5$ for all r. $Y|(D=1, R=r) \sim N(1,1)$ for all r and $Y|(D=0, R=r) \sim N(0,1)$ for all r.

Size2 Same as Size1 except that

$$P(D=1|R=r) = \mathbf{1}\{-2 \le r < 0\}\frac{(r+2)^2}{8} + \mathbf{1}\{0 \le r \le 2\}\left(1 - \frac{(r-2)^2}{8}\right).$$

In both DGPs, the propensity scores are continuous at the cut-off (i.e., Assumption 3 does not hold). Combined with FRD-validity (Assumptions 1 and 2), the distributions of the observables are also continuous at the cut-off, implying that these DGPs correspond to least favorable nulls in the context of our test. Size1 has a constant propensity score, while in Size2, the left- and right-derivatives of the propensity scores differ at the cut-off.

For each DGP, we generate random samples of four sizes: 1000, 2000, 4000, and 8000 observations. We specify $\mathcal{L}^* = \mathcal{L}_{coarse}$ with $Q = 15$.[12] For each simulation design, we conduct 1000 repetitions with $\bar{B} = 300$ bootstrap iterations. We consider three data-driven choices of bandwidths: Imbens and Kalyanaraman (2012, IK), Calonico, Cattaneo, and Titiunik (2014, CCT), and Arai and Ichimura (2016, AI). For each bandwidth, we impose undersmoothing by multiplying $n^{\frac{1}{5}-\frac{1}{c}}$ and the bandwidth, choosing $c = 4.5$.[13] In addition, we also consider the MSE-optimal robust bias correction (MSE-RBC) implementation (see Calonico, Cattaneo, and Farrell, 2018) and the coverage error rate-optimal (CER-RBC) implementation (see Calonico, Cattaneo, and Farrell, 2020). For the MSE-RBC bandwidth, we implement the test by estimating the conditional means via local quadratic regression using a bandwidth that is MSE-optimal for local linear regression (AI, IK, or CCT), see Calonico, Cattaneo, and Titiunik (2014, Remark 7). For the CER-RBC bandwidth, we multiply a MSE-optimal

[12]We note that our test exhibits similar results when Q is greater than 10.

[13]We run simulations for other choices of the under-smoothing constant $c \in [3,5)$; the results are similar.

bandwidth (AI, IK, or CCT) by the rule-of-thumb adjustment factor proposed in Calonico, Cattaneo, and Farrell (2020, section 4).

TABLE 1. Rejection Frequency at the 5% Level

DGP	n	US AI	US IK	US CCT	MSE-RBC AI	MSE-RBC IK	MSE-RBC CCT	CER-RBC AI	CER-RBC IK	CER-RBC CCT
Size1	1000	0.060	0.02	0.019	0.037	0.016	0.017	0.055	0.025	0.024
	2000	0.071	0.025	0.034	0.058	0.018	0.022	0.076	0.022	0.024
	4000	0.078	0.038	0.035	0.067	0.049	0.022	0.086	0.043	0.035
	8000	0.065	0.045	0.033	0.066	0.051	0.037	0.057	0.045	0.036
Size 2	1000	0.063	0.014	0.012	0.039	0.008	0.015	0.067	0.015	0.014
	2000	0.064	0.035	0.031	0.051	0.033	0.021	0.060	0.032	0.024
	4000	0.064	0.041	0.038	0.077	0.040	0.037	0.060	0.042	0.039
	8000	0.060	0.039	0.036	0.065	0.042	0.040	0.057	0.044	0.035

Table 1 summarizes the results at the 5% nominal level. For the full set of results at other significance levels, see Tables F.1 to F.3 in Appendix F of the Supplemental Material. The results show that the proposed test controls size well for each of the specified designs and the various bandwidth choices. Although our test statistic (which takes the supremum over a class of intervals) is different from those considered in Calonico, Cattaneo, and Farrell (2020), and Calonico, Cattaneo, and Farrell (2020), the CER-RBC and MSE-RBC implementations work well.

4.2. **Power properties.** To investigate the power properties, we consider the following four DGPs, Power1-Power4, in which the conditional distribution of Y_1 violates the local continuity condition in different ways.[14]

Power1 Let $R \sim N(0,1)$ truncated at -2 and 2. The propensity score is given by

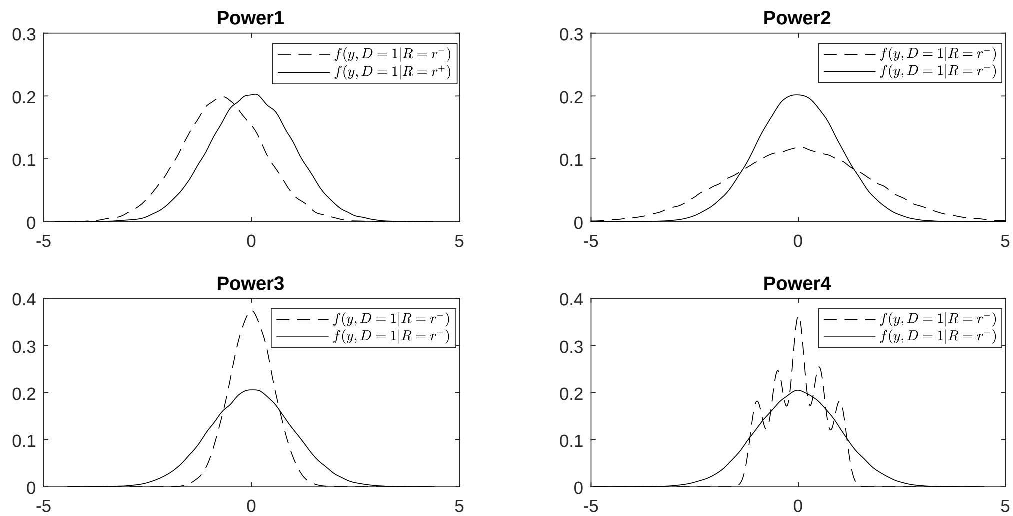
$$P(D=1|R=r) = \mathbf{1}\{-2 \le r < 0\} \max\{0, (r+2)^2/8 - 0.01\} \\ + \mathbf{1}\{0 \le r \le 2\} \min\{1, 1-(r-2)^2/8 + 0.01\}$$

Let $Y|(D=0, R=r) \sim N(0,1)$ for all $r \in [-2,2]$, and $Y|(D=1, R=r) \sim N(0,1)$ for all $r \in [0,2]$. Let $Y|(D=1, R=r) \sim N(-0.7,1)$ for all $r \in [-2,0)$.

Power2 Same as Power1 except that $Y|(D=1, R=r) \sim N(0, 1.675^2)$ for all $r \in [-2,0)$.

[14] Appendix F of the Supplemental Material provide examples where violation of the local monotonicity assumption or the local continuity assumption results in distributions of observables similar to those for Power1.

FIGURE 1. Potential Outcome Densities at the Cut-off



Power3 Same as Power1 except that $Y|(D=1, R=r) \sim N(0, 0.515^2)$ for all $r \in [-2, 0)$.

Power4 Same as Power1 except that $Y|(D=1, R=r) \sim \sum_{j=1}^{5} \omega_j N(\mu_j, 0.125^2)$ for all $r \in [-2, 0)$, where $\omega = (0.15, 0.2, 0.3, 0.2, 0.15)$ and $\mu = (-1, -0.5, 0, 0.5, 1)$.

Figure 1 plots the potential outcome density at the cut-off for each of Power1-Power4, in which the testable implication of Theorem 1 (i) is violated since the solid curves and the dashed curves intersect. Table 2 reports simulation results for the power properties of our test at the 5% level. Additional results are collected in Tables F.4 to F.6 in Appendix F. Overall, our test has good power in detecting deviations from the null under all choices of bandwidth. It is harder for our test to reject in Power4. From Figure 1, we see that the violation of the null in Power4 occurs abruptly with many peaks over narrow intervals, whereas in the other designs (e.g., Power1 and Power2) mild violation occurs over relatively wide intervals. This phenomenon is consistent with what has been noted in the literature: the Bierens (1982)-, and Andrews and Shi (2013)-type methods that we adopt in this paper are efficient in detecting the second type of violations.[15]

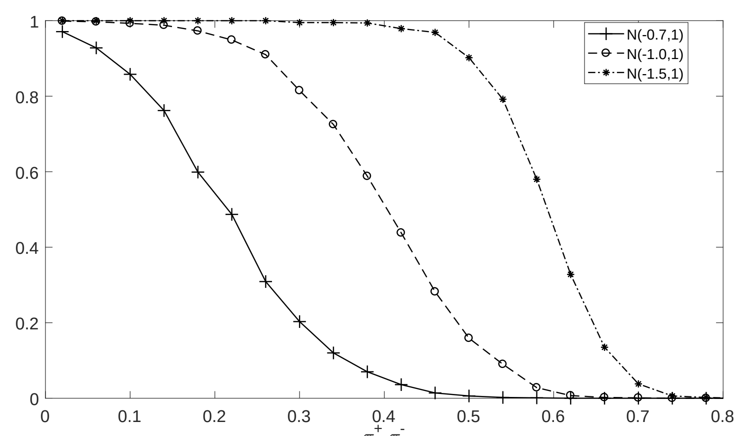
As the magnitude of the propensity score jump $\pi^+ - \pi^-$ becomes smaller, we expect that the inequalities of (2) and (3) become closer to binding. For instance, in the extreme case of

[15]See Chernozhukov, Lee, and Rosen (2013, footnote 10) for related discussion.

TABLE 2. Rejection Frequency at the 5% Level

DGP	n	US AI	US IK	US CCT	MSE-RBC AI	MSE-RBC IK	MSE-RBC CCT	CER-RBC AI	CER-RBC IK	CER-RBC CCT
	1000	0.215	0.174	0.111	0.103	0.081	0.050	0.225	0.143	0.090
Power1	2000	0.439	0.403	0.256	0.221	0.207	0.115	0.439	0.316	0.205
	4000	0.753	0.744	0.604	0.476	0.459	0.305	0.749	0.629	0.486
	8000	0.962	0.975	0.907	0.812	0.820	0.654	0.964	0.935	0.831
	1000	0.122	0.061	0.052	0.086	0.023	0.022	0.133	0.046	0.045
Power2	2000	0.271	0.194	0.140	0.135	0.099	0.052	0.266	0.156	0.097
	4000	0.554	0.511	0.342	0.293	0.246	0.142	0.560	0.399	0.248
	8000	0.885	0.888	0.732	0.624	0.622	0.391	0.889	0.793	0.598
	1000	0.164	0.123	0.078	0.106	0.063	0.027	0.159	0.107	0.061
Power3	2000	0.299	0.257	0.170	0.174	0.154	0.079	0.306	0.209	0.128
	4000	0.573	0.510	0.383	0.361	0.289	0.183	0.581	0.421	0.321
	8000	0.883	0.870	0.734	0.694	0.640	0.466	0.888	0.781	0.640
	1000	0.099	0.050	0.024	0.057	0.024	0.017	0.101	0.036	0.027
Power4	2000	0.172	0.123	0.060	0.118	0.060	0.042	0.175	0.092	0.074
	4000	0.264	0.268	0.144	0.181	0.144	0.079	0.265	0.201	0.138
	8000	0.550	0.540	0.326	0.341	0.326	0.201	0.545	0.438	0.283

FIGURE 2. Power and Propensity Jump Size



$\pi^+ - \pi^- = 0$, for a distribution satisfying the testable implication, inequalities (2) and (3) must hold with equality, i.e., the conditional distribution of $(Y, D)|R$ is continuous at the cut-off. This means a joint distribution of potential outcomes and selection type violating FRD-validity is more likely to violate the testable implications as the magnitude of the jump in the propensity score becomes smaller. In the opposite direction, the testable implication of Theorem 1 loses screening power when the FRD design is close to a sharp design.

We illustrate this point by modifying the propensity score of Power1 to

$$P(D=1|R=r) = \mathbf{1}\{-2 \le r < 0\} \max\{0, (r+2)^2/8 - d\} \\ + \mathbf{1}\{0 \le r \le 2\} \min\{1, 1-(r-2)^2/8 + d\}.$$

Here, $2d$ measures the jump size of the propensity score and $d = 0.01$ corresponds to the results of Power1. In addition to the specification $Y|(D=1, R=r) \sim N(-0.7, 1)$ for $r \in [-2, 0)$, we consider two additional specifications, $Y|(D=1, R=r) \sim N(-1, 1)$ and $Y|(D=1, R=r) \sim N(-1.5, 1)$ for $r \in [-2, 0)$, which lead to larger deviations from the null.

Figure 2 plots the rejection frequency as a function of $\pi^+ - \pi^- = 2d$ for each of the alternative distributions at the 5% level for a sample size of 8000 observations.[16] At each specification of $Y|(D=1, R=r)$ for $r \in [-2, 0)$, we see that the rejection frequency decreases as the jump size increases. As the jump size approaches one (the sharp design), the rejection frequency falls to zero because inequalities (2) and (3) are never violated in the sharp design. On the other hand, for a given jump size, a larger deviation from local continuity leads to a larger rejection frequency, as expected.

5. Applications

To illustrate that implementing our test can provide new insights for empirical practice, we assess FRD-validity in the designs studied in Angrist and Lavy (1999, AL hereafter), and Miller, Pinto, and Vera-Hernández (2013, MPV hereafter).

5.1. **Effect of class size on student performance.**

Israel has been implementing Maimonides' rule in public schools since 1969. The rule limits a class size to 40 students and so creates discontinuous changes in the average class size as the total enrollment exceeds multiples of 40 students. For example, a public school with 40 enrolled students in a grade can maintain one class, with a (average) class size of 40 students; another public school with 41 enrolled students has to offer two classes, and so the average class size drops discontinuously from 40 students to 20.5 students. Maimonides' rule

[16] Here we only report the results based on the under-smoothed IK bandwidth. Other choices produce similar results.

offers an example of FRD design since some schools in the data do not comply with the treatment assignment rule.[17]

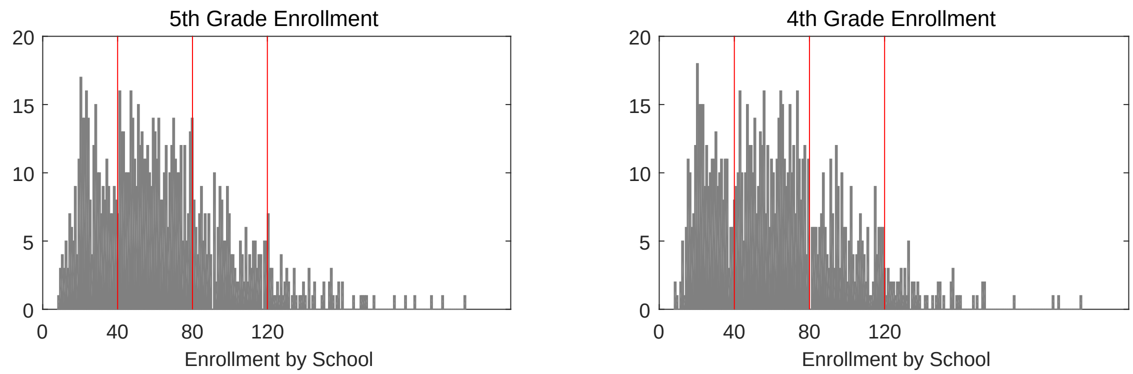
Recent empirical evidence suggests that the density of the running variable (enrollment) is discontinuous near some cut-offs (Otsu, Xu, and Matsushita (2013), and Angrist, Lavy, Leder-Luis, and Shany (2019)). Along with the argument of Lee (2008), and McCrary (2008), this evidence raises concerns about FRD-validity, but cannot be interpreted as direct evidence to refute local continuity or local monotonicity.

Who manipulates class size? As argued in AL, parents may selectively exploit Maimonides' rule by either (a) registering their children into schools with enrollments slightly above multiples of 40 students, hoping that their children will be placed in smaller classes, or (b) withdrawing children from those public schools with enrollments slightly below multiples of 40 students. In either case, we expect to observe discontinuities of the density of the running variable at the cut-offs, as we can observe most notably at the enrollment count of 40 students in Figure 3. Class size manipulation by parents can be a serious threat to the local continuity assumption if those parents who act according to (a) also more highly value a small class-size education, and are more concerned with their children's education. If children with such parents perform better than their peers, the potential outcome distributions of the students' test scores violate local continuity.

On the other hand, AL defend FRD-validity by arguing that manipulation of class size by parents is not likely. Concerning the possibility of (a), AL claim that: "*there is no way [for the parents] to know [exactly] whether a predicted enrollment of 41 [students] will not decline to 38 [students] by the time school starts, obviating the need for two small classes*". With respect to the possibility of (b), private elementary schooling is rare in Israel and withdrawing their children is not a feasible option for most parents. Angrist, Lavy, Leder-Luis, and Shany (2019) re-investigate Maimonides' rule and argue that the manipulation is operated mainly on the school board side, stating that: "*A recent memo from Israeli Ministry of Education (MOE) officials to school leaders admonishes headmasters against*

[17] We define the treatment as whether the school "splits" ($D = 1$) or does "not split" ($D = 0$) a cohort with an enrollment around the cut-off into smaller classes. Focusing on grade 4, with bandwidth equal to 3 and the cut-off at 80 students as an example, we first restrict the sample to classes if their schools' grade 4 enrollment is $R \in \{78, 79, 80\} \cup \{81, 82, 83\}$ students. Then we assign $D = 1$ to a class if its school has "three classes" and assign $D = 0$ to a class if its school has "two classes". In the data, there are schools that have an enrollment within $\{78, 79, 80\} \cup \{81, 82, 83\}$ students but that have either one or more than three classes. They are very rare (about 0.2% of the total observations), and we exclude these observations from our analysis.

FIGURE 3. Histograms for Enrollments by Schools: Panel A of Figure 6 in Angrist, Lavy, Leder-Luis, and Shany (2019)



attempts to increase staffing ratios through enrollment manipulation. In particular, schools are warned not to move students between grades or to enroll those who are overseas so as to produce an additional class." This type of manipulation can lead to a density jump like that observed in Figure 3, but is not necessarily a serious threat to FRD-validity depending on the school board's incentives to manipulate. If the main motivation of manipulation is to increase their budget (an increasing function of the number of classes), as argued in Angrist, Lavy, Leder-Luis, and Shany (2019), and if the distributions of the students' potential outcomes in those schools where boards manipulate enrollment are the same as those in schools where boards do not manipulate, any manipulation around the cut-off is independent of the students' unobserved talents. Then, FRD-validity can hold even when the density of the running variable is discontinuous at the cut-offs.

Test Results. The testable implication assessed by our test focuses on the joint distribution of the observed outcomes and treatment status, in contrast to the density continuity approach that focuses only on the marginal distribution of the running variable. Hence, our test can provide new empirical evidence that can contribute to the dispute about the FRD-validity of Maimonides' rule, reviewed above.

We apply the test proposed in Section 3 for each of the four outcome variables (grade 4 math and verbal test scores, and grade 5 math and verbal test scores) by treating the three cut-offs of 40, 80, and 120 students, separately. We consider the bandwidths ($h_+ = h_- = 3$ and $h_+ = h_- = 5$)

used in AL, as well as the three data-driven bandwidth choices (AI, IK and CCT).[18] We also report p-values using the RBC bandwidth choices based on CCT with MSE-Optimal and CER-Optimal criteria, respectively. We set the trimming constant to $\xi = 0.00999$, as described in Algorithm 1 of Section 3.[19]

Table 3 displays the p-values of the tests. For all the cases considered, we do not reject the null hypothesis at a 10% significance level. The results are robust to the choice of bandwidths and the choice of trimming constants (see Tables G.2 to G.4 in Appendix G of the Supplemental Material). Despite the fact that the density of the running variable appears to be discontinuous at the cut-off, "no rejection" by our test suggests empirical support for the argument of "manipulation by the school board"–the type of manipulation that is relatively innocuous for AL's identification strategy. As discussed in Section 4.2 and illustrated in Figure 2, it is, however, important to acknowledge that the statistical power of our test might be limited by the large jumps in the propensity score that occur in this application, ranging from 0.3 to 0.7 (see Table G.1 in the Supplemental Material).

5.2. **Colombia's Subsidized Regime.** MPV study the impact of "Colombia's Régimen Subsidiado (SR)," a publicly financed insurance program targeted at poor households, on financial risk protection, service use, and health outcomes. SR subsidizes eligible Colombians to purchase insurance from private and government-approved insurers. Program eligibility is determined by a threshold rule based on a continuous index called Sistema de Identificacion de Beneficiarios (SISBEN) ranging from 0 to 100 (with 0 being the most impoverished, and those below a cut-off being eligible). SISBEN is constructed by a proxy means-test using fourteen different measurements of a household's well-being. It is, however, well known that the original SISBEN index used to assign the actual program eligibility was manipulated by either households or the administering authority (see MPV and the references therein for details). To circumvent this issue of manipulation, MPV simulate their own SISBEN index for each household using a collection of survey data from independent sources. MPV then estimate a cut-off of the simulated SISBEN scores in each region by maximizing the performance of in-sample prediction for the actual program take-up. Using these estimated cut-offs, MPV estimate the compliers' effects of SR on 33 outcome variables in four categories: (i) risk

[18] See Table G.11 in Appendix G of the Supplemental Material for the obtained bandwidths and the number of observations therein.

[19] We try different choices for the trimming constant $\xi \in \{0.0316, 0.1706, 0.5\}$ and obtain similar results.

TABLE 3. Testing Results for Israeli School Data: p-values, $\xi = 0.00999$

	3	5	AI	IK	CCT	MSE-RBC	CER-RBC
g4math							
Cut-off 40	0.986	0.934	0.767	0.978	0.968	0.964	0.975
Cut-off 80	0.909	0.865	0.715	0.944	0.888	0.771	0.957
Cut-off 120	0.443	0.702	0.665	0.604	0.568	0.613	0.639
g4verb							
Cut-off 40	0.928	0.627	0.465	0.648	0.529	0.564	0.463
Cut-off 80	0.911	0.883	0.185	0.906	0.720	0.284	0.842
Cut-off 120	0.935	0.683	0.474	0.730	0.186	0.228	0.143
g5math							
Cut-off 40	0.876	0.282	0.488	0.631	0.609	0.901	0.265
Cut-off 80	0.516	0.446	0.930	0.482	0.765	0.808	0.726
Cut-off 120	0.939	0.827	0.626	0.883	0.838	0.842	0.772
g5verb							
Cut-off 40	0.594	0.893	0.953	0.906	0.938	0.955	0.962
Cut-off 80	0.510	0.692	0.504	0.525	0.929	0.953	0.973
Cut-off 120	0.696	0.811	0.601	0.699	0.781	0.739	0.745

protection, consumption smoothing, and portfolio choice, (ii) medical care use, (iii) health status, and (iv) behavior distortions; see Table G.5 of the Supplemental Material and Table 1 of MPV for details.

Although the density of the simulated SISBEN score passes the continuity test (see MPV's online Appendix C), it does not necessarily imply FRD-validity, e.g., the conditional distributions of the potential outcomes given the simulated SISBEN score may not be continuous at the cut-off.

For each of the 33 outcome variables, we implement our test using MPV's simulated SISBEN score as the running variable and the actual program enrollment as the treatment status. We consider the three bandwidths ($h_+ = h_- = 2, 3,$ and 4) used in MPV as well as the three data-driven bandwidth choices (AI, IK and CCT).[20] We use the same set of trimming constants ξ as in the AL application and find that the results are insensitive to a choice of ξ. We find robust evidence to reject the testable implications of FRD-validity for the following three outcome variables: "household education spending," "total spending on food," and "total monthly expenditure." Their p-values are

[20]See Table G.12 in Appendix G of the Supplemental Material for the obtained bandwidths and the number of observations contained therein.

reported in Table 4 (results for all other outcome variables and other choices of ξ are collected in Tables G.5-G.8 in Appendix G of the Supplemental Material).

TABLE 4. Testing Results for Columbia's SR Data: p-values ($\xi = 0.00999$)

Outcome variables	MPV Bandwidths			Other Bandwidth Choices				
	2	3	4	AI	IK	CCT	MSE RBC	CER RBC
Household education spending	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Total spending on food	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Total monthly expenditure	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

A few remarks are in order. First, the three outcome variables giving the robust rejections all belong to the first category: "risk protection, consumption smoothing, and portfolio choice." For other outcome variables, we do not find evidence against FRD-validity. The low p-values for these three outcomes remain significant even when we take into account multiple-testing of a group of outcome variables with family-wise error rate (FWER) control.[21]

Second, it is possible to figure out which observations cause the rejection of FRD-validity. Take the choice of $\mathcal{L}^* = \mathcal{L}_{coarse}$ with $Q = 15$ and bandwidth $h^+ = h^- = 2$, and focus on the outcome variable "total spending on food" as an example. The supremum in the test statistic is achieved at $d = 0$ and $[y, y'] = [0.5948, 0.6527]$. Figure 4 draws the kernel smoothed (pseudo) densities of the normalized outcome variable, where the blue curve should be underneath the red curve under FRD-validity. The two density curves in the top-right panel indeed cross in this interval. The histograms and Table 5 show that there are 45 observations with $D = 0, R \in (r_0 - 2, r_0)$ and $Y \in [0.5948, 0.6527]$, which are about 6.73% of all observations with $R \in (r_0 - 2, r_0)$ and $D = 0$. On the other hand, there are 31 observations with $D = 0, R \in [r_0, r_0 + 2)$ and $Y \in [0.5948, 0.6527]$, which is about 2.06% of all observations with $R \in [r_0, r_0 + 2)$ and $D = 0$. See Figures G.1, G.2 and Tables G.13, G.14 in the Supplemental Material for similar analysis on the other two outcome variables in Table 4.

Third, we also condition on each of the six regions in Colombia when implementing the test. The results are collected in Table G.9 in the Supplemental Material. We obtain strong rejections

[21]The results shown in Table G.5 of Appendix G of the Supplemental Material imply that, for the first category of 10 outcome variables, the multiple testing procedure of Holm (1979) concludes that the joint null hypothesis of FRD-validity holding for the 10 outcomes is rejected for the control of FWER at 1%. With all the outcomes (33 hypotheses), the joint null hypothesis is rejected for the control of FWER at 5%.

TABLE 5. Obs. in the maximizer interval ($h^+ = h^- = 2$): Total spending on food

	# of observations		
Subsample of	All	$\{0.5948 \leq Y \leq 0.6527\}$	Ratio
$\{0 \leq R < h^+\} \cap \{D = 0\}$ ($\mathbf{N} \cup \mathbf{C}$)	1502	31	2.06%
$\{h^- < R < 0\} \cap \{D = 0\}$ ($\mathbf{N}$)	669	45	6.73%

in the "Atlantica", "Oriental", "Central", and "Bogota" regions, and no rejection in "Pacifico" and "Territorios Nacionales". Taking into account the relative sample sizes across the regions (Table G.10 of the Supplemental Material), the Bogota sample seems to drive the test results of Table 4. Notice that the magnitude of the propensity score jump for the Bogota sample is relatively small compared with the samples in the regions giving no-rejections (see Figure 5). This observation is in line with Figure 2 and the discussion of Section 4.

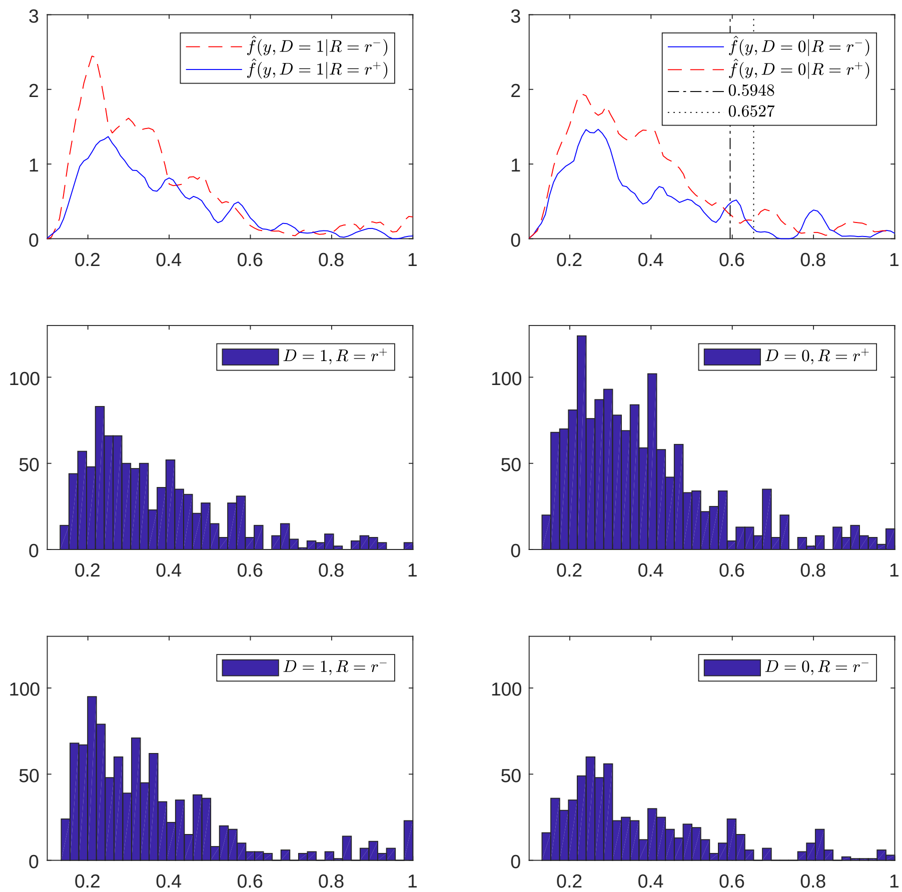
There are several possible reasons why FRD-validity fails in this application. First, violation of local continuity may arise as a byproduct of estimating the cut-off using the simulated SISBEN score. For instance, if there is some household characteristic that is *not* included in the construction of the simulated SISBEN score but has strong predictive power for program enrollment, the estimated cut-off may pick up a value of the simulated SISBEN score across which the distribution of the excluded characteristic differs most. If the distribution of household consumption variables depend on such an excluded characteristic, the result is a violation of local continuity. Second, there could be other *unobserved* programs using the same SISBEN index with similar cut-offs. If such programs significantly affect a household's budget, we can expect the distribution of potential household consumption to be quite different on each of the two sides of the cut-off, again leading to a violation of local continuity.[22]

6. CONCLUSION

In this paper we propose a specification test for the key identifying conditions in fuzzy regression design. We characterize the set of sharp testable implications for FRD-validity and propose an asymptotically valid test for it. Our approach makes use of not only the information conveyed by the

[22]MPV suggest that the second channel is less likely to be the cause of rejection of FRD-validity for the relevant three outcome variables. See Table 2 in MPV for evidence that the enrollment rates for other programs do not change across the estimated cut-offs.

FIGURE 4. Estimated complier's outcome density: Total spending on food



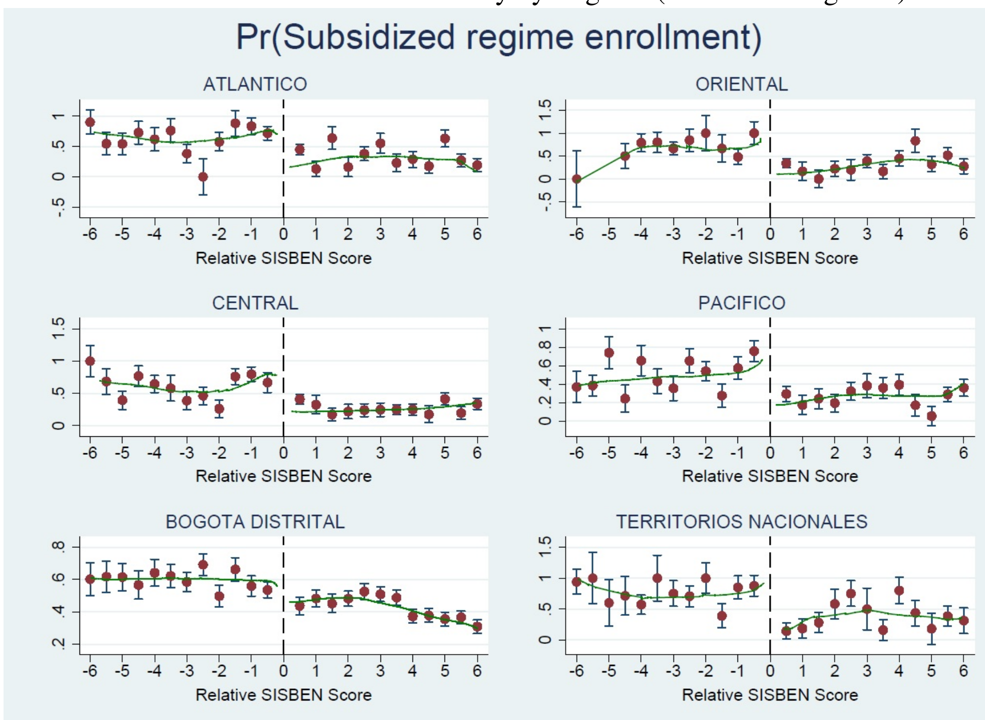
running variable but also that conveyed by the outcome and treatment status. As illustrated in our empirical applications, our specification test provides empirical evidence for or against FRD-validity, which is overlooked if only the continuity of the running variable's density at the cut-off is assessed.

APPENDIX A. CALCULATING THE TEST STATISTICS

We describe how to compute the proposed test statistic. Let $m_{P,d}(\ell,r) = \mathbb{E}_P[g_\ell(Y)D^d(1-D)^{1-d}|R=r]$ and $m_{P,d,+}(\ell) = \lim_{r\downarrow r_0} m_{P,d}(\ell,r)$ and $m_{P,d,-}(\ell) = \lim_{r\uparrow r_0} m_{P,d}(\ell,r)$ for $d=1,0$, then we can estimate $\nu_{P,1}(\ell)$ and $\nu_{P,0}(\ell)$ respectively by equation (9), which we restate below:

$$\hat{\nu}_1(\ell) = \hat{m}_{1,-}(\ell) - \hat{m}_{1,+}(\ell), \quad \hat{\nu}_0(\ell) = \hat{m}_{0,+}(\ell) - \hat{m}_{0,-}(\ell),$$

FIGURE 5. Enrollment Probability by Regions (from MPV Figure 2)



where the right hand side terms $\hat{m}_{d,\star}(\ell)$, for $d = 1, 0$ and $\star = +, -$, are local linear estimators. They can be constructed by the intercept estimates $\hat{a}_{d,+}(\ell)$ and $\hat{a}_{d,-}(\ell)$ in the local regressions,

$$\begin{aligned}
&(\hat{a}_{d,+}(\ell), \hat{b}_{d,+}(\ell)) \\
&\quad = \underset{a,b}{\operatorname{argmin}} \frac{1}{nh_+} \sum_{i=1}^{n} 1\{R_i \geq r_0\} \cdot K\left(\frac{R_i - r_0}{h_+}\right) \left[g_\ell(Y_i) D_i^d (1 - D_i)^{1-d} - a - b \cdot \left(\frac{R_i - r_0}{h_+}\right)\right]^2, \\
&(\hat{a}_{d,-}(\ell), \hat{b}_{d,-}(\ell)) \\
&\quad = \underset{a,b}{\operatorname{argmin}} \frac{1}{nh_-} \sum_{i=1}^{n} 1\{R_i < r_0\} \cdot K\left(\frac{R_i - r_0}{h_-}\right) \left[g_\ell(Y_i) D_i^d (1 - D_i)^{1-d} - a - b \cdot \left(\frac{R_i - r_0}{h_-}\right)\right]^2,
\end{aligned}$$

where $K(\cdot)$ is a kernel function and (h_+, h_-) are the bandwidths specified above and below the cut-off, respectively. In particular, we express $h_+ = c_+ h$ and $h_- = c_- h$, with (c_+, c_-) being positive constants and h being a sequence converging to zero as $n \to \infty$. For simplicity of analysis and implementation, we specify the bandwidths h_+ and h_- to be the same over $\{g_\ell : \ell \in \mathcal{L}\}$.

We can write the local linear estimators in the following form: for $d = 1, 0$ and $\star = +, -$

$$\hat{m}_{d,\star}(\ell) = \sum_{i=1}^{n} w_{n,i}^{\star} \cdot g_\ell(Y_i) D_i^d (1 - D_i)^{1-d}, \tag{11}$$

where the weights are defined as

$$w_{n,i}^{+} = \frac{1}{nh_+} \cdot \frac{1\{R_i \geq r_0\} \cdot K(\frac{R_i - r_0}{h_+}) \left[\hat{\vartheta}_2^+ - \hat{\vartheta}_1^+ \cdot \left(\frac{R_i - r_0}{h_+}\right)\right]}{\hat{\vartheta}_2^+ \hat{\vartheta}_0^+ - (\hat{\vartheta}_1^+)^2},$$

$$w_{n,i}^{-} = \frac{1}{nh_-} \cdot \frac{1\{R_i < r_0\} \cdot K(\frac{R_i - r_0}{h_-}) \left[\hat{\vartheta}_2^- - \hat{\vartheta}_1^- \cdot \left(\frac{R_i - r_0}{h_-}\right)\right]}{\hat{\vartheta}_2^- \hat{\vartheta}_0^- - (\hat{\vartheta}_1^-)^2}.$$

and for $j = 0, 1, 2$,

$$\hat{\vartheta}_j^+ = \frac{1}{nh_+} \sum_{i=1}^{n} 1\{R_i \geq r_0\} \cdot K\left(\frac{R_i - r_0}{h_+}\right) \left(\frac{R_i - r_0}{h_+}\right)^j,$$

$$\hat{\vartheta}_j^- = \frac{1}{nh_-} \sum_{i=1}^{n} 1\{R_i < r_0\} \cdot K\left(\frac{R_i - r_0}{h_-}\right) \left(\frac{R_i - r_0}{h_-}\right)^j.$$

REFERENCES

ANDREWS, D. W. K., AND X. SHI (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81(2), 609–666.

ANDREWS, D. W. K., AND X. SHI (2014): "Nonparametric inference based on conditional moment inequalities," *Journal of Econometrics*, 179(1), 31–45.

ANDREWS, D. W. K., AND G. SOARES (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78(1), 119–157.

ANGRIST, J. D., AND V. LAVY (1999): "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *Quarterly Journal of Economics*, 114(2), 533–575.

ANGRIST, J. D., V. LAVY, J. LEDER-LUIS, AND A. SHANY (2019): "Maimonides Rule Redux," *American Economic Review Insights*, 1, 309–324.

ARAI, Y., Y.-C. HSU, T. KITAGAWA, I. MOURIFIÉ, AND Y. WAN (2020): "Supplement to "Testing Identifying Assumptions in Fuzzy Regression Discontinuity Designs"," *Quantitative Economics Supplemental Material*.

ARAI, Y., AND H. ICHIMURA (2016): "Optimal bandwidth selection for the fuzzy regression discontinuity estimator," *Economics Letters*, 141, 103–106.

ARAI, Y., AND T. KITAGAWA (2016): "A specification test in fuzzy regression discontinuity design," *unpublished manuscript*.

ARMSTRONG, T. B., AND H. P. CHAN (2016): "Multiscale adaptive inference on conditional moment inequalities," *Journal of Econometrics*, 194(1), 24–43.

BALKE, A., AND J. PEARL (1997): "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 92(439), 1171–1176.

BARRETT, G. F., AND S. G. DONALD (2003): "Consistent Tests for Stochastic Dominance," *Econometrica*, 71, 71–104.

BERTANHA, M., AND G. W. IMBENS (2020): "External Validity in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 38, 593–612.

BIERENS, H. J. (1982): "Consistent model specification tests," *Journal of Econometrics*, 20(1), 105–134.

BUGNI, F. A., AND I. A. CANAY (2018): "Testing Continuity of a Density via g order statistics in the Regression Discontinuity Design," *Working Paper*.

CALONICO, S., M. D. CATTANEO, AND M. H. FARRELL (2018): "On the effect of bias estimation on coverage accuracy in nonparametric inference," *Journal of the American Statistical Association*, 113(522), 767–779.

——— (2020): "Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs," *The Econometrics Journal*, 23(2), 192–210.

CALONICO, S., M. D. CATTANEO, AND R. TITIUNIK (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82(6), 2295–2326.

CANAY, I. A., AND V. KAMAT (2018): "Approximate permutation tests and induced order statistics in the regression discontinuity design," *Review of Economic Studies*, 85, 1577–1608.

CARNEIRO, P., AND R. GINJA (2014): "Long-term impacts of compensatory preschool on health and behavior: Evidence from Head Start," *American Economic Journal: Economic Policy*, 6(4), 135–173.

CATTANEO, M. D., AND J. C. ESCANCIANO (eds.) (2017): *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing.

CATTANEO, M. D., M. JANSSON, AND X. MA (2020): "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, 115(531), 1449–1455.

CATTANEO, M. D., L. KEELE, R. TITIUNIK, AND G. VAZQUEZ-BARE (2016): "Interpreting regression discontinuity designs with multiple cutoffs," *Journal of Politics*, 78(4), 1229–1248.

CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): "Intersection bounds: estimation and inference," *Econometrica*, 81(2), 667–737.

CHETVERIKOV, D. (2018): "Adaptive Tests of Conditional Moment Inequalities," *Econometric Theory*, 34(1), 186–227.

DELL, M. (2010): "The persistent effects of Peru's mining *mita*," *Econometrica*, 78(6), 1863–1903.

DONALD, S. G., AND Y.-C. HSU (2016): "Improving the power of tests of stochastic dominance," *Econometric Reviews*, 35(4), 553–585.

DONG, Y. (2018): "Alternative assumptions to identify LATE in fuzzy regression discontinuity designs," *Oxford Bulletin of Economics and Statistics*, 80(5), 1020–1027.

DONG, Y., AND A. LEWBEL (2015): "Identifying the effect of changing the policy threshold in regression discontinuity models," *Review of Economics and Statistics*, 97(5), 1081–1092.

EUGSTER, B., R. LALIVE, A. STEINHAUER, AND J. ZWEIMÜLLER (2017): "Culture, work, attitudes, and job search: evidence from the Swiss language border," *Journal of the European Economic Association*, 15(5), 1056–1100.

FRANDSEN, B. R., M. FRÖLICH, AND B. MELLY (2012): "Quantile treatment effects in the regression discontinuity design," *Journal of Econometrics*, 168(2), 382–395.

GERARD, F., M. ROKKANEN, AND C. ROTHE (2020): "Bounds on Treatment Effects in Regression Discontinuity Designs with a Manipulated Running Variable," *Quantitative Economics*, 11, 839–870.

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.

HANSEN, B. E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64(2), 413–430.

HANSEN, P. R. (2005): "A test for superior predictive ability," *Journal of Business & Economic Statistics*, 23, 365–380.

HASTINGS, J. S., C. A. NEILSON, AND S. D. ZIMMERMAN (2014): "Are some degrees worth more than others? Evidence from college admission cutoffs in Chile," *NBER working paper*, (19241).

HECKMAN, J. J., AND E. VYTLACIL (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73(3), 669–738.

HOLM, S. (1979): "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, 6(2), 65–70.

HSU, Y.-C., I. MOURIFIÉ, AND Y. WAN (2016): "Testing identifying assumptions in regression discontinuity designs," *unpublished manuscript*.

IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62(2), 467–475.

IMBENS, G. W., AND K. KALYANARAMAN (2012): "Optimal bandwidth choice for the regression discontinuity estimator," *The Review of Economic Studies*, 79(3), 933–959.

IMBENS, G. W., AND T. LEMIEUX (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142(2), 615–635.

IMBENS, G. W., AND D. B. RUBIN (1997): "Estimating outcome distributions for compliers in instrumental variables models," *The Review of Economic Studies*, 64(4), 555–574.

KEELE, L. J., AND R. TITIUNIK (2015): "Geographic boundaries as regression discontinuities," *Political Analysis*, 23, 127–155.

KIRKEBOEN, L. J., E. LEUVEN, AND M. MOGSTAD (2016): "Field of study, earnings, and self-selection," *Quarterly Journal of Economics*, 131(3), 1057–1111.

KITAGAWA, T. (2015): "A test for instrument validity," *Econometrica*, 83(5), 2043–2063.

KOLESÁR, M., AND C. ROTHE (2018): "Inference in regression discontinuity designs with a discrete running variable," *American Economic Review*, 108(8), 2277–2304.

LEE, D. S. (2008): "Randomized experiments from non-random selection in US House elections," *Journal of Econometrics*, 142(2), 675–697.

LEE, D. S., AND T. LEMIEUX (2010): "Regression discontinuity designs in economics," *Journal of Economic Literature*, 48(2), 281–355.

LINTON, O., K. SONG, AND Y.-J. WHANG (2010): "An improved bootstrap test of stochastic dominance," *Journal of Econometrics*, 154(2), 186–202.

McCrary, J. (2008): "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 142(2), 698–714.

Miller, G., D. Pinto, and M. Vera-Hernández (2013): "Risk protection, service use, and health outcomes under Colombia's health insurance program for the poor," *American Economic Journal: Applied Economics*, 5(4), 61–91.

Mourifié, I., and Y. Wan (2017): "Testing local average treatment effect assumptions," *Review of Economics and Statistics*, 99(2), 305–313.

Otsu, T., K.-L. Xu, and Y. Matsushita (2013): "Estimation and inference of discontinuity in density," *Journal of Business & Economic Statistics*, 31(4), 507–524.

Rubin, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of Educational Psychology*, 66(5), 688.

Stinebrickner, R., and T. R. Stinebrickner (2014): "A major in science? Initial beliefs and final outcomes for college major and dropout," *Review of Economic Studies*, 81(1), 426–472.

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.

Zafar, B. (2011): "How do college students form expectations?," *Journal of Labor Economics*, 29(2), 301–348.