# Testing Monotonicity of Conditional Treatment Effects Under Regression Discontinuity Designs

**Yu-Chin Hsu**[†] **Shu Shen**[‡]

This version: Dec. 20, 2019

## Abstract

Researchers are often interested in the relationship between treatment effects and observed individual heterogeneity. This paper proposes the first nonparametric monotonicity test for conditional treatment effects under the popular regression discontinuity framework. The proposed tests examine whether the average treatment effect or the local average treatment effect has a monotonic relationship with some of the observed individual characteristics. We show consistency and asymptotic uniform size control of the proposed tests. We apply the test to study treatment effect heterogneity in seminal RD studies in Pop-Eleches and Urquiola (2013) and Lindo et al. (2010).

**Keywords:** average treatment effect, local average treatment effect, regression discontinuity, regression monotonicity

[†] Yu-Chin Hsu, Institute of Economics, Academia Sinica; Department of Finance, National Central University; Department of Economics, National Chengchi University. E-mail: ychsu@econ.sinica.edu.tw.

[‡] Shu Shen, Department of Economics, University of California, Davis. E-mail: shushen@ucdavis.edu.

**Acknowledgement:** Yu-Chin Hsu gratefully acknowledges research support from the Ministry of Science and Technology of Taiwan (MOST107-2410-H-001-034-MY3) and the Career Development Award of Academia Sinica, Taiwan. Shu Shen gratefully acknowledges research support from the Hellman Fellows Award. All errors are the authors'.

# 1 Introduction

In program evaluation, researchers are often interested in the whole picture of a treatment effect that is beyond the overall population average. Estimators and tests of treatment effect heterogeneity (e.g., Heckman et al., 1998; Abadie, 2002; Hotz et al., 2005; Firpo, 2007; Crump et al., 2008; Wager and Athey, 2018, among many others) therefore play an important role in the literature. Researchers use these methods to quantify treatment effects for different groups of individuals, to look for relationships between the effects and observed factors, to understand how a policy intervention can affect tails of an outcome distribution, or to design extensions of the analyzed treatment to other populations.

Treatment effect heterogeneity analysis is also important in program evaluation studies that use the regression discontinuity (RD) design, which has become very popular in applied microeconomics, following pioneering works of Angrist and Lavy (1999), Black (1999), and van der Klaauw (2002). In RD models, the probability of an individual receiving a policy treatment changes discontinuously with an underlying variable, often referred to as the running variable. By comparing the values of the response outcome above and below the discontinuity point of the running variable, researchers identify effects for individuals at the margin of policy treatment nonparametrically. In this paper, we propose the first nonparametric RD monotonicity test for examining whether a conditional average treatment effect (CATE) identified under a sharp RD design or a conditional local average treatment effect (CLATE) identified under a fuzzy RD design has a monotonic relationship with some of the observed individual characteristics.

The proposed test is important to the RD literature because applied economists are often interested in testing for such a relationship. For example, Ito (2015) studies the treatment effect of an electricity rebate program in California and is interested in knowing whether the effect on electricity consumption increases with average temperature or decreases with household income. Carneiro et al. (2015) investigate the impact of increased maternity leave on children's long-term outcomes in Norway and examine how the effect changes with family characteristics such as mothers' education, distance to grandparents, and income. Barone et al. (2015) study the impact of being exposed to slanted information on decision making using a quasi-natural experiment in Italy. They

find that switching to digital TV, which has tenfold more programs, decreased the share of votes received by Berlusconi's coalition, and the effect was stronger in towns with older and less educated voters.

Despite their popularity, tools used in the applied literature to test for monotonic relationships between treatment effects and observed individual heterogeneity are quite informal. One popular methodology, as is adopted, for example, in Ito (2015) and Barone et al. (2015), is to add interaction terms between control variables for individual heterogeneity and the dummy variable indicating whether an observation passes the discontinuity point of the running variable. The method would conclude that the treatment effect on average increases (decreases) with a certain observable if the corresponding interaction term is positive (negative) and statistically significant. Another popular approach is to split the data by values of additional controls and carry out subsample RD analysis. For example, Carneiro et al. (2015) group individuals by quartiles of mothers' months of unpaid leave and log income and examine whether subsample treatment effect estimates change monotonically from low to high quartiles. Unfortunately, neither method described above is ideal. The interaction term method is parametric and subject to model misspecification. The subsampling method often uses ad-hoc grouping criteria and may result in information loss. In contrast, our proposed test is nonparametric, consistent, and have asymptotic uniform size control.

Our test also contributes to the regression monotonicity literature in statistics and econometrics. Existing nonparametric regression monotonicity tests (e.g., Ghosal et al., 2000; Hall and Heckman, 2000; Chetverikov, 2019; Hsu et al., 2019), to the best of our knowledge, consider only situations involving interior estimation. Since the RD treatment effect is estimated through nonparametric boundary estimation, none of the existing monotonicity tests could be applied to the setup. Our test is hence the first that is compatible with nonparametric boundary estimation.

To construct our test, we first formulate the null hypothesis of regression monotonicity as a conditional moment inequality that conditions on both the running variable of the RD model and some other controls. We then use instrumental functions to transform the moment inequality into a series of conditional moment inequalities that condition only on the running variable and build our test statistic upon the latter. Critical values are

constructed through multiplier bootstrap. The proposed nonparametric RD monotonicity test has uniform size control over a broad set of data generating processes and is robust to weak identification of the CLATE under fuzzy RD. The latter is important because even when the identification of a local average treatment effect (LATE) is strong, part of the CLATE function could be weakly identified due to first-stage heterogeneity. In addition, our proposed test statistic does not involve nonparametric derivative estimation and is of order $(nh)^{-1/2}$, the same rate of convergence as the classic RD treatment effect estimators, regardless of the dimension of the conditioning variable in CATE or CLATE.

The instrumental function approach adopted in the paper is related to Andrews and Shi (2013, 2014) and Hsu et al. (2019), who propose to reduce the dimension of the conditioning set in the conditional moment (in)equalities by transforming the outcome variable with a series of countably many instrumental functions. They also show that such a transformation does not result in loss of information. Our test is closest to Hsu et al. (2019), who extend the instrumental function approach in Andrews and Shi (2013, 2014) to test for generalized regression monotonicity. However, as is discussed earlier, the general method developed in Hsu et al. (2019) cannot be applied to the RD framework because their test is not compatible with nonparametric boundary estimators.

This paper is also related to the large literature on regression discontinuity, especially some recent developments that study treatment effect heterogeneity. Relevant papers include Frandsen et al. (2012) and Shen and Zhang (2016) for distributional RD analysis, Dong and Lewbel (2015) and Angrist and Rokkanen (2015) for extrapolating RD effects away from the cut-off, Bertanha (2016) and Cattaneo et al. (2016) for analyzing heterogeneous treatment effect when the RD design has multiple cut-offs, Bertanha and Imbens (2014) for the examining the external validity of LATE under the fuzzy RD design, Hsu and Shen (2019) for testing whether a treatment effect is heterogeneous among individuals with different observed characteristics, Frölich and Huber (2019) and Calonico et al. (2019) for increasing the precision of treatment effect estimation, and Frölich and Huber (2019) for restoring RD validity when a covariate distribution is allowed to change abruptly at the RD cut-off.

We apply the tests to two empirical examples. The first example studies the impact of attending a more selective high school in Romania using data of Pop-Eleches and

Urquiola (2013). Mean analysis in Pop-Eleches and Urquiola (2013) finds that going to a better high school does not significantly change the probability of a student taking the Baccalaureate exam but improves grade averages among those who take the exam. Interestingly, our proposed nonparametric test reveals that the effect on the exam-taking rate is monotonically larger if the more selective school has higher-achieving peers, indicating that the insignificant mean effect found in Pop-Eleches and Urquiola (2013) may come from the cancelation of positive and negative treatment effects among different schools.

The second example studies the effect of being placed on academic probation at the end of the first year in college using data in Lindo et al. (2010). Lindo et al. (2010) find that the discouragement effect of academic probation "is greater for students that performed relatively better in high school (above the median of students entering the university)" by comparing students with above-median high school grades to those with below-median high school grades. As is discussed earlier, such a subsampling approach uses an arbitrary data-splitting rule and often does not adjust for multiple testing. Our proposed nonparametric test treats high school grade percentile as a continuous random variable and does not find evidence supporting the monotonicity argument stated above.

The paper is organized as follows. Section 3 sets up the RD model and proposes the benchmark monotonicity test for the general fuzzy RD case. Section 3 discusses the asymptotic property of the proposed test. Section 4 extends the benchmark test to the special case of sharp RD. Section 5 examines the small sample performance of the proposed test using Monte Carlo simulations, and Section 6 carries out the two empirical applications. Proofs of all lemmas and theorems are provided in the Appendix.

## 2 Model Set-up

Let $Y$ denote the outcome of interest and $T$ the binary treatment status of an individual. $T = 0$ if an individual does not take the treatment and $T = 1$ if he/she does. Use $Y(0)$ and $Y(1)$ to denote potential outcomes when $T = 0$ and $T = 1$, respectively. The observed outcome is $Y = T \cdot Y(1) + (1 - T) \cdot Y(0)$. Whether an individual receives treatment or not depends at least partially on an underlying running variable $Z$. A policy intervention encourages an individual to receive treatment if $Z$ is larger than or equal to

some known cut-off value $c$. Let $T(1)$ and $T(0)$ be the potential treatment decisions of an individual depending on whether he/she is encouraged or not. The observed treatment status is then $T = T(1) \cdot 1(Z \geq c) + T(0) \cdot 1(Z < c)$. The model is said to follow a fuzzy RD design if the treatment decision $T$ is a probabilistic function of $Z$. Individuals with $T(1) = T(0) = 1$, $T(1) = T(0) = 0$, $T(1) - T(0) = 1$, and $T(0) - T(1) = 1$ are called always-takers, never-takers, compliers, and defiers, respectively. When the treatment decision $T$ is a deterministic function of $Z$ such that $T = 1(Z \geq c)$, everyone in the population is a complier and the model is said to follow the sharp RD design. Let $X$ be a set of covariates with compact and convex support $\mathcal{X} \subset R^{d_x}$.

Let $P$ be the underlying distribution of $(Z, T, X, Y)$ and use $E_P$ to denote expectation under $P$. Under the general case of fuzzy RD, the CLATE defined as

$$CLATE(x) = E_P[Y(1) - Y(0)|X = x, Z = c, T(1) - T(0) = 1]$$

captures the average treatment effect of compliers at the RD cut-off and with specific values of $X$. The dependence of $CLATE(x)$ on $P$ is suppressed for notational simplicity.

As is discussed in the introduction, researchers are often interested in treatment effect heterogeneity and examining whether the $CLATE(x)$ has a monotonic relationship with the value of $X$. Let $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ be the subset of $X$ values that are of interest. The following assumption collects the identifying conditions of $CLATE(x)$ for $x \in \tilde{\mathcal{X}}$.

**Assumption 2.1** *For any $\delta > 0$, define $\mathcal{N}_{\delta,z}(c) = \{z : |z-c| \leq \delta\}$ as the $\delta$-neighborhood of $Z$ around the cut-off value $c$. For a running variable $Z$ continuously distributed in $\mathcal{N}_{\delta,z}(c)$ for some $\delta > 0$, assume that*

*(i)* *$E_P[Y(t)|T(1)-T(0) = 1, X = x, Z = z]$ and $E_P[Y(t)|T(1) = T(0) = t', X = x, Z = z]$ are continuous in $z$ on $\tilde{\mathcal{X}} \times \mathcal{N}_{\delta,z}(c)$ for $t, t' \in \{0, 1\}$;*

*(ii)* *$E_P[T(1) - T(0) = 1|X = x, Z = z]$ and $E_P[T(1) = T(0) = t|X = x, Z = z]$ are continuous in $z$ on $\tilde{\mathcal{X}} \times \mathcal{N}_{\delta,z}(c)$ for $t \in \{0, 1\}$;*

*(iii)* *$T(1) \geq T(0)$;*

*(iv)* *$E_P[T(1) - T(0)|X = x, Z = c] > 0$ for all $x \in \tilde{\mathcal{X}}$.*

Assumptions 2.1 is standard in RD (c.f. Dong and Lewbel, 2015) except that (i), (ii) and (iv) require a stronger version of the classic conditions that do not condition on $X$. These stronger conditions are necessary as we want to identify the CLATE given values of $X$. Under Assumption 2.1, it is straightforward to show that

$$\begin{aligned} CLATE(x) &= \frac{E_P[(Y(1)-Y(0))(T(1)-T(0))|X=x,Z=c]}{E_P[T(1)-T(0)|X=x,Z=c]} \\ &= \frac{\lim_{z\searrow c} E_P[Y|X=x,Z=z] - \lim_{z\nearrow c} E_P[Y|X=x,Z=z]}{\lim_{z\searrow c} E_P[T|X=x,Z=z] - \lim_{z\nearrow c} E_P[T|X=x,Z=z]} \end{aligned} \tag{2.1}$$

for all $x \in \tilde{\mathcal{X}}$.

When the treatment status is a deterministic function of the running variable and the model is of sharp RD design, researchers are interested in the CATE defined as

$$CATE(x) = E_P[Y(1)-Y(0)|X=x,Z=c].$$

Since everyone is a complier under sharp RD, the identification restriction reduce to the following condition.

**Assumption 2.2** *For a running variable $Z$ continuously distributed in $\mathcal{N}_{\delta,z}(c)$ for some $\delta > 0$, $E_P[Y(t)|X=x,Z=z]$ is continuous in $z$ on $\tilde{\mathcal{X}} \times \mathcal{N}_{\delta,z}(c)$ for $t = 0, 1$.*

Under Assumption 2.2, the CATE is identified as

$$CATE(x) = \lim_{z\searrow c} E_P[Y|X=x,Z=z] - \lim_{z\nearrow c} E_P[Y|X=x,Z=z] \tag{2.2}$$

for all $x \in \tilde{\mathcal{X}}$.

# 3 Benchmark Test

This section considers a simple benchmark test with a continuous single-dimensional conditioning variable $X$. General cases of the test will be discussed in Section 4. Furthermore, for notational simplicity, we assume without loss of generality that $\tilde{\mathcal{X}} = \mathcal{X}$, or that the subset of covariate values of interest in hypothesis testing is equal to the full domain of $X$. We also assume without loss of generality that the domain $\mathcal{X}$ is equal to the unit interval $[0, 1]$, as all compact and convex sets can be rescaled to the unit interval.

### 3.1 Null Hypothesis: Fuzzy RD Design

First we consider the case of fuzzy RD design. The null and alternative hypotheses of interest can be formalized as

$$H_{0,FRD}: \ CLATE(x) \text{ is non-decreasing in } x \text{ on } \mathcal{X}; \tag{3.1}$$
$$H_{1,FRD}: \ H_{0,FRD} \text{ does not hold.}$$

In our first empirical application in Section 6.1, $X$ is the peer quality of the more selective school measured by the average entrance score of peers, and $CLATE(.)$ is the average RD effect of attending a better high school on the probability of a student taking the college entrance exam conditional on the peer quality of the more selective school. In the second empirical application in Section 6.2, we have data from a sharp RD design and therefore $CLATE(.) = LATE(.)$. In that example, $X$ is a student's high school grade percentile, and $CLATE(.)$ is the average RD effect of being placed on academic probation at the end of the freshmen year on the probability of a student leaving college conditional on the student's high school grade. We will discuss the sharp RD design specifically in Section 3.5.

To test the null $H_{0,FRD}$, a direct approach is to take the partial derivative of $CLATE(x)$ with respect to $x$ and examine the sign of the derivative function for all $x \in \mathcal{X}$. This approach would require both the numerator and the denominator of $CLATE(x)$ to be differentiable with respect to $x$. It also requires the use of nonparametric derivative estimation, which has a slow convergence rate. In this paper, we take an alternative route following the idea of Hsu et al. (2019). We transform the null of (3.1) to conditional moment inequalities that do not involve derivatives, as is illustrated in the following Lemma.

**Lemma 3.1** *If $\lambda(x)$ is a continuous function on $\mathcal{X} = [0,1]$ and $h(x)$ is a weighting function satisfying $0 < h(x) \leq M < \infty$, then the following two statements are equivalent:*

*(i) $\lambda(x_1) \geq \lambda(x_2)$ whenever $x_1 > x_2$ and $x_1, x_2 \in \mathcal{X} = [0,1]$.*

*(ii) For any* $q \in \{2, 3, \cdots\}$,

$$\frac{\int_{x_1}^{x_1+1/q} \lambda(x) \cdot h(x)dx}{\int_{x_1}^{x_1+1/q} h(x)dx} \geq \frac{\int_{x_2}^{x_2+1/q} \lambda(x) \cdot h(x)dx}{\int_{x_2}^{x_2+1/q} h(x)dx} \text{ or equivalently} \tag{3.2}$$

$$\int_{x_2}^{x_2+1/q} \lambda(x) \cdot h(x)dx \cdot \int_{x_1}^{x_1+1/q} h(x)dx - \int_{x_1}^{x_1+1/q} \lambda(x) \cdot h(x)dx \cdot \int_{x_2}^{x_2+1/q} h(x)dx \leq 0. \tag{3.3}$$

*whenever* $x_1 > x_2$ *and* $x_1, x_2 \in \{0, 1/q, 2/q, \cdots, (q-1)/q\}$.

Lemma 3.1 states that monotonicity of any continuous function stated in (i) can be re-formulated as countably many moment inequalities given in (ii) and the transformation does not result in loss of information. Intuitively, because the left-hand side and right-hand side of (3.2) are simply weighted averages of $\lambda(x)$ over $[x_1, x_1 + q^{-1}]$ and $[x_2, x_2 + q^{-1}]$, the monotonicity condition in (i) implies the inequalities in (ii) directly. When $q = 2$, the inequality in (3.2) is simply $\int_{1/2}^{1} \lambda(x) \cdot h(x)dx / \int_{1/2}^{1} h(x)dx \geq \int_{0}^{1/2} \lambda(x) \cdot h(x)dx / \int_{0}^{1/2} h(x)dx$, reflecting the fact that the weighted average of $\lambda(x)$ is weakly higher when averaged over $[1/2, 1]$ than over $[0, 1/2]$. When $q = 3$, the result in (3.2) rewrites to three pairs of comparisons among the weighted averages calculated over $[0, 1/3]$, $[1/3, 2/3]$ and $[2/3, 1]$. To extract all information in (i), it is then necessary to consider in (ii) all possible intervals $[x_1, x_1+q^{-1}]$ and $[x_2, x_2+q^{-1}]$ constructed by combinations of $(x_1, x_2, q)$ for $q = 2, 3, \cdots$.

The use of the weighting function $h(x)$ in Lemma 3.1 is illustrated in the following. Let $\lambda(x) = CLATE(x)$ and assume the function is continuous in $x$. The statement in (i) reduces to the null hypothesis $H_{0,FRD}$. When $h(x)$ is defined as $h(x) = E_P[T(1) - T(0)|X = x, Z = c] \cdot f_{X|Z}(x|c)$ with $f_{X|Z}(x|z)$ denoting the conditional density of $X = x$ given $Z = z$, the inequalities in (3.3) in the statement in (ii) reduce to

$$\begin{aligned}
&E_P[1(X \in [x_2, x_2 + 1/q])E_P[(Y(1) - Y(0))(T(1) - T(0))|X, Z = c]|Z = c] \\
&\times E_P[1(X \in [x_1, x_1 + 1/q])E_P[T(1) - T(0)|X, Z = c]|Z = c] \\
-&E_P[1(X \in [x_1, x_1 + 1/q])E_P[(Y(1) - Y(0))(T(1) - T(0))|X, Z = c]|Z = c] \\
&\times E_P[1(X \in [x_2, x_2 + 1/q])E_P[T(1) - T(0)|X, Z = c]|Z = c] \leq 0.
\end{aligned} \tag{3.4}$$

We see that the weighting function $h(x)$ in Lemma 3.1 is designed to accommodate $\lambda(x)$ functions in the form of fractions so that the transformed moment inequalities would be fraction-free so that the test statistics to be proposed based on the transformed moment inequalities could get around the presence of random denominators. Lemma 3.1 is therefore more general than Lemma 3.1 of Hsu et al. (2019), where $\lambda(x)$ and $h(x)$ are specifically set to $E[Y|X=x]$ and $f_X(x)$, respectively.

Under Assumption 2.1 and with continuity of $f_{X|Z}(x|z)$, the inequlities in (3.4) can be further identified as

$$\begin{aligned}
&\left(\lim_{z\searrow c} E_P[g_{x_2,q}(X)Y|Z=z] - \lim_{z\nearrow c} E_P[g_{x_2,q}(X)Y|Z=z]\right) \\
&\times \left(\lim_{z\searrow c} E_P[g_{x_1,q}(X)T|Z=z] - \lim_{z\nearrow c} E_P[g_{x_1,q}(X)T|Z=z]\right) \\
&- \left(\lim_{z\searrow c} E_P[g_{x_1,q}(X)Y|Z=z] - \lim_{z\nearrow c} E_P[g_{x_1,q}(X)Y|Z=z]\right) \\
&\times \left(\lim_{z\searrow c} E_P[g_{x_2,q}(X)T|Z=z] - \lim_{z\nearrow c} E_P[g_{x_2,q}(X)T|Z=z]\right) \leq 0, \qquad (3.5)
\end{aligned}$$

where $g_{x_t,q}(X) \equiv 1\left(X \in [x_t, x_t + 1/q]\right)$, for $t = 1, 2$, $x_1, x_2 \in \{0, 1/q, 2/q, \cdots, (q-1)/q\}$, and $q \in \{2, 3, ...\}$. Let $\ell = (x_1, x_2, q)$ and

$$\mathcal{L} \equiv \left\{(x_1, x_2, q) : \ (x_1, x_2) \in \{0, 1/q, 2/q, \cdots, (q-1)/q\}^2, x_1 > x_2, \text{ for } q = 2, 3 \cdots\right\} \qquad (3.6)$$

be its support. We use $\mu_P(\ell)$ to denote the left-hand side of (3.5) for fixed values of $x_1$, $x_2$, and $q$. Then writing the inequalities in (3.5) for all $x_1, x_2 \in \{0, 1/q, 2/q, \cdots, (q-1)/q\}$ and $q \in \{2, 3, ...\}$ is equivalent to writing $\mu_P(\ell) \leq 0$ for all $\ell \in \mathcal{L}$. In the next, we formalize the above described transformation results for the monotonicity of $CLATE(.)$.

**Assumption 3.1** *Assume that:*

1. *$f_{X|Z}(x|z)$ is continuous and uniformly bounded in $x$ and $z$ on $\mathcal{X} \times \mathcal{N}_{\delta,z}(c)$.*
2. *The conditional expectations in Assumptions 2.1.(i) and (ii) are continuous in $x$ so that $CLATE(x)$ is continuous in $x$ on $\mathcal{X}$.*

Assumption 3.1.(i) is a direct implication of the "no precise control over the running variable" rule introduced by Lee and Lemieux (2010) for pre-determined controls and is

a well-accepted concept in RD applications. See Hsu and Shen (2019) for a detailed discussion. Assumption 3.1.(ii) is required to apply the transformation result in Lemma 3.1 to the monotonicity test of interest.

**Lemma 3.2** *Under Assumptions 2.1 and 3.1, the null hypothesis $H_{0,FRD}$ is equivalent to*

$$H_{0,FRD}^{T} : \mu_P(\ell) \equiv \rho_P^{(2)}(\ell)\varrho_P^{(1)}(\ell) - \rho_P^{(1)}(\ell)\varrho_P^{(2)}(\ell) \leq 0, \quad \text{for all } \ell \in \mathcal{L}, \tag{3.7}$$

*where* $m_{P,+}^{(\kappa)}(\ell) = \lim_{z \searrow c} E_P[g_{x_\kappa,q}(X)Y|Z=z]$, $m_{P,-}^{(\kappa)}(\ell) = \lim_{z \nearrow c} E_P[g_{x_\kappa,q}(X)Y|Z=z]$, $q_{P,+}^{(\kappa)}(\ell) = \lim_{z \searrow c} E_P[g_{x_\kappa,q}(X)T|Z=z]$, $q_{P,-}^{(\kappa)}(\ell) = \lim_{z \nearrow c} E_P[g_{x_\kappa,q}(X)T|Z=z]$, $\rho_P^{(\kappa)}(\ell) = m_{P,+}^{(\kappa)}(\ell) - m_{P,-}^{(\kappa)}(\ell)$, $\varrho_P^{(\kappa)}(\ell) = q_{P,+}^{(\kappa)}(\ell) - q_{P,-}^{(\kappa)}(\ell)$, *for both* $\kappa = 1, 2$, *and* $\mathcal{L}$ *is defined in* (3.6).

A formal proof of Lemma 3.2 is given in the Appendix. In the next section, we will construct the test statistic using the transformed moment inequalities described in the Lemma. As is seen from the form of the inequalities, the proposed test statistic will not involve plug-in estimators of $CLATE(.)$, which can have non-classical asymptotic properties under weak identification due to the random denominator problem (see, for example, Feir et al., 2016). This is especially important for our proposed test to have good small sample behavior because even in cases where unconditional effects such as $LATE$ are strongly identified, first stage heterogeneity can easily result in weak identification of $CLATE(.)$ in some part of the function. Use of plug-in estimators of $CLATE(.)$ would, therefore, result in size distortion in finite samples and shall be avoided.

## 3.2 Test Statistic: Fuzzy RD Design

Let $\{Z_i, T_i, X_i, Y_i\}_{i=1}^n$ be a sample of size $n$ randomly drawn from the underlying distribution of $(Z, T, X, Y)$. Recall that $\ell = (x_1, x_2, q)$ is used to index the transformed moment inequalities, and $\ell \in \mathcal{L}$. For $\kappa = 1, 2$, let $\hat{m}_{n,+}^{(\kappa)}(\ell)$, $\hat{m}_{n,-}^{(\kappa)}(\ell)$, $\hat{q}_{n,+}^{(\kappa)}(\ell)$ and $\hat{q}_{n,-}^{(\kappa)}(\ell)$ be the local linear estimators of $m_{P,+}^{(\kappa)}(\ell)$, $m_{P,-}^{(\kappa)}(\ell)$, $q_{P,+}^{(\kappa)}(\ell)$ and $q_{P,-}^{(\kappa)}(\ell)$ defined in Lemma 3.2. Let $K(.)$ be the kernel function and $h$ the bandwidth. Local linear estimators $\hat{m}_{n,+}^{(\kappa)}(\ell)$, $\hat{m}_{n,-}^{(\kappa)}(\ell)$, $\hat{q}_{n,+}^{(\kappa)}(\ell)$ and $\hat{q}_{n,-}^{(\kappa)}(\ell)$ are the constant terms of the following minimization prob-

lems, respectively.

$$\left(\hat{m}_{n,+}^{(\kappa)}(\ell), \hat{b}_{n,m+}^{(\kappa)}(\ell)\right) = \arg\min_{a,b} \sum_{Z_i \geq c}^{n} K\left(\frac{Z_i - c}{h}\right) \left[g_{x_\kappa,q}(X_i)Y_i - a - b(Z_i - c)\right]^2,$$

$$\left(\hat{m}_{n,-}^{(\kappa)}(\ell), \hat{b}_{n,m-}^{(\kappa)}(\ell)\right) = \arg\min_{a,b} \sum_{Z_i < c}^{n} K\left(\frac{Z_i - c}{h}\right) \left[g_{x_\kappa,q}(X_i)Y_i - a - b(Z_i - c)\right]^2,$$

$$\left(\hat{q}_{n,+}^{(\kappa)}(\ell), \hat{b}_{n,q+}^{(\kappa)}(\ell)\right) = \arg\min_{a,b} \sum_{Z_i \geq c}^{n} K\left(\frac{Z_i - c}{h}\right) \left[g_{x_\kappa,q}(X_i)T_i - a - b(Z_i - c)\right]^2,$$

$$\left(\hat{q}_{n,-}^{(\kappa)}(\ell), \hat{b}_{n,q-}^{(\kappa)}(\ell)\right) = \arg\min_{a,b} \sum_{Z_i < c}^{n} K\left(\frac{Z_i - c}{h}\right) \left[g_{x_\kappa,q}(X_i)T_i - a - b(Z_i - c)\right]^2.$$

Following Fan and Gijbels (1992), define

$$S_{n,l}^{+} = \sum_{i=1}^{n} 1(Z_i \geq c) K\left(\frac{Z_i - c}{h}\right)(Z_i - c)^l, \;\; S_{n,l}^{-} = \sum_{i=1}^{n} 1(Z_i < c) K\left(\frac{Z_i - c}{h}\right)(Z_i - c)^l,$$

$$\mathsf{w}_{ni}^{+} = 1\left(Z_i \geq c\right) K\left(\frac{Z_i - c}{h}\right)\left[S_{n,2}^{+} - S_{n,1}^{+}(Z_i - c)\right] / \left[S_{n,0}^{+}S_{n,2}^{+} - S_{n,1}^{+}S_{n,1}^{+}\right],$$

$$\mathsf{w}_{ni}^{-} = 1\left(Z_i < c\right) K\left(\frac{Z_i - c}{h}\right)\left[S_{n,2}^{-} - S_{n,1}^{-}(Z_i - c)\right] / \left[S_{n,0}^{-}S_{n,2}^{-} - S_{n,1}^{-}S_{n,1}^{-}\right].$$

for $l = 0, 1, 2$, and re-write the local linear estimators as

$$\hat{m}_{n,+}^{(\kappa)}(\ell) = \sum_{i=1}^{n} \mathsf{w}_{ni}^{+} \cdot m^{(\kappa)}(Y_i, X_i, \ell), \quad \hat{m}_{n,-}^{(\kappa)}(\ell) = \sum_{i=1}^{n} \mathsf{w}_{ni}^{-} \cdot m^{(\kappa)}(Y_i, X_i, \ell),$$

$$\hat{q}_{n,+}^{(\kappa)}(\ell) = \sum_{i=1}^{n} \mathsf{w}_{ni}^{+} \cdot q^{(\kappa)}(T_i, X_i, \ell), \quad \hat{q}_{n,-}^{(\kappa)}(\ell) = \sum_{i=1}^{n} \mathsf{w}_{ni}^{-} \cdot q^{(\kappa)}(T_i, X_i, \ell),$$

where $m^{(\kappa)}(Y_i, X_i, \ell) = g_{x_\kappa,q}(X_i)Y_i$, and $q^{(\kappa)}(T_i, X_i, \ell) = g_{x_\kappa,q}(X_i)T_i$, for $\ell \in \mathcal{L}$ and $\kappa = 1, 2$.

Let $\hat{\rho}_n^{(\kappa)}(\ell) = \hat{m}_{n,+}^{(\kappa)}(\ell) - \hat{m}_{n,-}^{(\kappa)}(\ell)$ and $\hat{\varrho}_n^{(\kappa)}(\ell) = \hat{q}_{n,+}^{(\kappa)}(\ell) - \hat{q}_{n,-}^{(\kappa)}(\ell)$ for $\ell \in \mathcal{L}$ and $\kappa = 1, 2$. Then

$$\hat{\mu}_n(\ell) = \hat{\rho}_n^{(2)}(\ell)\hat{\varrho}_n^{(1)}(\ell) - \hat{\rho}_n^{(1)}(\ell)\hat{\varrho}_n^{(2)}(\ell) \tag{3.8}$$

is an estimator of the moment function $\mu_P(\ell)$. We show in the Appendix in Lemma B.2 that $\hat{\mu}_n(.)$ is uniformly consistent. Let $\hat{\sigma}_{\mu,n}^2(.)$ be a uniformly consistent estimator for the variance of $\hat{\mu}_n(.)$, and $\hat{\sigma}_{\mu,n}^2(.)$ will be formally defined in the next section. Let $\epsilon$ be some small positive number, and $\ell_0 = (1/2, 0, 2)$. To test the null hypothesis $H_{0,FRD}$ described in the previous section, we use the Kolmogorov-Smirnov type statistic

$$\widehat{T}_{n,FRD} = \sup_{\ell \in \mathcal{L}} \sqrt{nh} \frac{\hat{\mu}_n(\ell)}{\hat{\sigma}_{\mu,\epsilon}(\ell)} \tag{3.9}$$

where $\hat{\sigma}^2_{\mu,\epsilon}(\ell) = \max\left\{\hat{\sigma}^2_{\mu,n}(\ell), \epsilon \cdot \hat{\sigma}^2_{\mu,n}(\ell_0)\right\}$ manually bounds the variance estimator away from zero. The tuning parameter $\epsilon$ is set to 0.005 in the simulation and empirical sections of the paper. Under regularity conditions, the null distribution of the test statistic is bounded by a known distribution when $H_{0,FRD}$ is true. The test statistic diverges if the monotonicity condition fails.

### 3.3 Decision Rule and Test Properties: Fuzzy RD Design

We introduce two simulated critical values for constructing decision rules of the proposed test. The critical value based on the traditional least favorable condition (LFC) method is straightforward but could be conservative under certain data generating processes (DGPs). To improve the power of the proposed test under such situations, we also consider the generalized moment selection (GMS) method introduced by Andrews and Soares (2010) and Andrews and Shi (2013, 2014, 2017).[1]

Let $\phi^{(\kappa)}_{\rho,ni}(.) = \sqrt{nh}(\mathsf{w}^+_{ni}(m^{(\kappa)}(Y_i, X_i, .) - m^{(\kappa)}_{P,+}(.)) - \mathsf{w}^-_{ni}(m^{(\kappa)}(Y_i, X_i, .) - m^{(\kappa)}_{P,-}(.)))$ and $\phi^{(\kappa)}_{\varrho,ni}(.) = \sqrt{nh}(\mathsf{w}^+_{ni}(q^{(\kappa)}(T_i, X_i, .) - q^{(\kappa)}_{P,+}(.)) - \mathsf{w}^-_{ni}(q^{(\kappa)}(T_i, X_i, .) - q^{(\kappa)}_{P,-}(.)))$. First, we notice that the moment estimator $\hat{\mu}_n(.)$ defined in the previous section has an influence function representation in the sense that

$$\begin{aligned}
&\sqrt{nh}(\hat{\mu}_n(\ell) - \mu_P(\ell)) \\
=&\sum_{i=1}^{n} \varrho^{(1)}_P(\ell) \cdot \phi^{(2)}_{\rho,ni}(\ell) + \rho^{(2)}_P(\ell) \cdot \phi^{(1)}_{\varrho,ni}(\ell) - \varrho^{(2)}_P(\ell) \cdot \phi^{(1)}_{\rho,ni}(\ell) - \rho^{(1)}_P(\ell) \cdot \phi^{(2)}_{\varrho,ni}(\ell) + o_p(1) \\
\equiv&\sum_{i=1}^{n} \phi_{\mu,ni}(\ell) + o_p(1),
\end{aligned}$$

where $o_p(1)$ is uniform over all $\ell \in \mathcal{L}$ under suitable conditions stated in Assumptions B.1-B.4 in the Appendix.

Replacing the population parameters in $\phi_{\mu,ni}(\ell)$ with their kernel estimation counterparts, we can define a consistent estimator of the control function. Let

$$\hat{\phi}_{\mu,ni}(\ell) \equiv \hat{\varrho}^{(1)}_n(\ell) \cdot \hat{\phi}^{(2)}_{\rho,ni}(\ell) + \hat{\rho}^{(2)}_n(\ell) \cdot \hat{\phi}^{(1)}_{\varrho,ni}(\ell) - \hat{\varrho}^{(2)}_n(\ell) \cdot \hat{\phi}^{(1)}_{\rho,ni}(\ell) - \hat{\rho}^{(1)}_n(\ell) \cdot \hat{\phi}^{(2)}_{\varrho,ni}(\ell), \tag{3.10}$$

[1]The recentering method proposed by Hansen (2005) and Donald and Hsu (2016), as well as the contact set approach proposed in Linton et al. (2010) could also be adopted to improve the power of our monotonicity tests.

where $\hat{\phi}^{(\kappa)}_{\rho,ni}(.)$ and $\hat{\phi}^{(\kappa)}_{\varrho,ni}(.)$ are defined by replacing $m^{(\kappa)}_{P,+}(.)$, $m^{(\kappa)}_{P,-}(.)$, $q^{(\kappa)}_{P,+}(.)$, and $q^{(\kappa)}_{P,-}(.)$ in $\phi^{(\kappa)}_{\rho,ni}(.)$ and $\phi^{(\kappa)}_{\varrho,ni}(.)$ with their estimators defined in the previous section. Let $\hat{\sigma}^2_{\mu,n}(\ell) = \sum_{i=1}^n \hat{\phi}^2_{\mu,ni}(\ell)$ be the variance estimator of $\hat{\mu}_n(\ell)$, for all $\ell \in \mathcal{L}$. Lemma B.4 proves uniform consistency of $\hat{\sigma}^2_{\mu,n}(.)$ as well as the estimator $\hat{\sigma}^2_{\mu,\epsilon}(.)$ defined in equation (3.9).

Let $\{U_i : 1 \leq i \leq n\}$ be a sequence of i.i.d. random variables that is independent of the sample path of $\{(Y_i, X_i, Z_i, T_i) : 1 \leq i \leq n\}$; $E[U_i] = 0$, $E[U_i^2] = 1$, and $E[|U_i|^4] < M$ for some $M > 0$. Under suitable conditions formulated in Assumptions B.1-B.4 in the Appendix, the simulated process

$$\widehat{\Phi}^u_{\mu,n}(.) = \sum_{i=1}^n U_i \cdot \hat{\phi}_{\mu,ni}(.) \tag{3.11}$$

would converge to the same limit as $\sqrt{nh}(\hat{\mu}_n(.) - \mu_P(.))$.

For significance level $\alpha < 1/2$, we define the LFC critical value as

$$\hat{c}^{\eta,LFC}_{n,FRD}(\alpha) = \sup\left\{ q \Big| P^u\Big( \sup_{\ell\in\mathcal{L}} \frac{\widehat{\Phi}^u_{\mu,n}(\ell)}{\hat{\sigma}_{\mu,\epsilon}(\ell)} \leq q \Big) \leq 1 - \alpha + \eta \right\} + \eta, \tag{3.12}$$

where $P^u$ is the multiplier probability measure. The LFC critical value is therefore the $(1-\alpha+\eta)$-th quantile of the simulated distribution of $\sup_{\ell\in\mathcal{L}} \frac{\widehat{\Phi}^u_{\mu,n}(\ell)}{\hat{\sigma}_{\mu,\epsilon}(\ell)}$ plus a small positive constant $\eta$, and the null hypothesis $H_{0,FRD}$ is rejected if $\hat{T}_{n,FRD} > \hat{c}^{\eta,LFC}_{n,FRD}(\alpha)$. Note that the small positive constant $\eta$ is required for the proposed test to have uniform size control. It could be set to zero if researchers only desire pointwise size control of the proposed test (c.f. Andrews and Guggenberger, 2009). Following Andrews and Shi (2013, 2014), we set $\eta = 10^{-6}$ in the rest of the paper. Similar to Andrews and Shi (2013, 2014), we also find in simulations and empirical applications that including this small $\eta$ value hardly makes any difference in practice.

The testing procedure based on the LFC critical value could be conservative, as the null hypothesis $H_{0,FRD}$ is one-sided. Alternatively, one could adopt the GMS method to construct the simulated critical value. For the significance level $\alpha < 1/2$, define the GMS critical value as

$$\hat{c}^{\eta,GMS}_{n,FRD}(\alpha) = \sup\left\{ q \Big| P^u\left( \sup_{\ell\in\mathcal{L}} \left( \frac{\widehat{\Phi}^u_{\mu,n}(\ell)}{\hat{\sigma}_{\mu,\epsilon}(\ell)} + \hat{\psi}_\mu(\ell) \right) \leq q \right) \leq 1 - \alpha + \eta \right\} + \eta, \tag{3.13}$$

$$\hat{\psi}_\mu(\ell) = -B_n \cdot 1\left( \sqrt{nh} \cdot \frac{\hat{\mu}_n(\ell)}{\hat{\sigma}_{\mu,\epsilon}(\ell)} < -a_n \right),$$

where $a_n$ and $B_n$ are two diverging sequences of non-negative numbers satisfying the conditions that $\lim_{n\to\infty} a_n/\sqrt{nh} = 0$, $\lim_{n\to\infty} B_n/a_n = 0$, and that $B_n$ is non-decreasing. Following Andrews and Shi (2013, 2014), we set $a_n = (0.3\ln(n))^{1/2}$ and $B_n = (0.4\ln(n)/\ln\ln(n))^{1/2}$ in the rest of the paper. Again, the decision rule is to reject the null when the test statistic is larger than the critical value. Compared to the LFC critical value, the GMS critical value uses the $\hat{\psi}_\mu(\ell)$ term to suppress the influence of the part of the simulated process corresponding to the negative moment functions on the critical value, and hence can improve the power of the proposed test.

Under proper assumptions on the smoothness of the underlying distribution of $(Z, T, X, Y)$, the kernel function $K(.)$, the bandwidth $h$, and regularity conditions for multiplier bootstrap, we formally show in the Appendix that the proposed test based on either the LFC or the GMS critical value achieves test consistency as well as uniform asymptotic size control. The uniform size control property holds over a compact subset of covariance kernels similar to the results in Theorem 2(a) of Andrews and Shi (2013). The benchmark test based on the LFC critical value is at most infinitesimally conservative asymptotically when there exists a constant $c$ such that $CLATE(x) = c$ for all $x \in \mathcal{X}$, while that based on the GMS critical value is at most infinitesimally conservative asymptotically when there exists a strictly positive measure of $x$ on which $CLATE(x) = c$ for some constant $c$. Detailed assumptions (Assumptions B.1-B.4) as well as theorems (Theorems B.1 and B.2) for uniform size control as well as test consistency are given in the Appendix.

### 3.4 Implementation Procedure: Fuzzy RD Design

In this section we summarize the implementation procedure of the proposed benchmark test. As discussed earlier, the kernel function $K(.)$ and the bandwidth $h$ for nonparametric estimation are regularized in Assumption B.2. We recommend using the triangular kernel function (i.e., $K(x) = |x| \cdot 1(|x| < 1)$), but the uniform kernel (i.e., $K(x) = \frac{1}{2} \cdot 1(|x| < 1)$) popular in empirical studies is also compatible with the assumption.

Assumption B.2 also requires undersmoothing of the kernel bandwidth to get rid of the nuisance bias term in the asymptotic limits of the random processes used to construct the test statistic. We suggest undersmoothing the robust RD bandwidth introduced

in Calonico et al. (2014) (CCT). Let $h = h_{CCT} \times n^{1/5-1/k}$, where $h_{CCT}$ is the CCT bandwidth and $k < 5$ is an under-smoothing parameter. In the simulation and empirical sections of the paper, we report testing results with different $k$ choices to examine the robustness of these results with respect to bandwidth choice.

Given the kernel function $K(.)$ and the bandwidth $h$ discussed above, we can carry out the proposed test with the following procedure.

1. Given the bandwidth $h$, trim the dataset such that all observations have the running variable lying between $c - h$ and $c + h$, where $c$ is the RD cut-off to speed up computation in the following steps.

2. Let $[\underline{x}, \bar{x}]$ be the support of $X$ in the trimmed dataset. Set $\delta_x = \bar{x} - \underline{x}$. For each $\ell = (x_1, x_2, q)$ such that $(x_1, x_2) \in \{\underline{x}, \underline{x} + \delta_x/q, \underline{x} + \delta_x \cdot 2/q, \cdots, \underline{x} + \delta_x \cdot (q-1)/q\}^2$, $x_1 > x_2$, and $q = 2, 3 \cdots Q$, estimate the moment function $\mu_P(\ell)$ and the influence function $\phi_{\mu,ni}(\ell)$ with their kernel estimators $\hat{\mu}_n(\ell)$ defined in (3.8) and $\hat{\phi}_{\mu,ni}$ defined in (3.10).

3. Construct the test statistic following (3.9) as the supreme of the standardized $\hat{\mu}_n(\ell)$ over all possible $\ell$ combinations used in Step 2.

4. Set a random seed and then draw random variables $U_1, U_2, ..., U_n \sim i.i.d.\ N(0,1)$ (or from other distributions that satisfy Assumption B.3) independent from the sample path. Simulate the process $\hat{\Phi}^u_{\mu,n}$ defined in (3.11). Based on the pre-determined significance level $\alpha$, calculate the LFC critical defined in (3.12) or the GMS ciritical value defined in (3.13).

5. Reject the null hypothesis if the test statistic constructed in Step 3 is larger than the critical value calculated in Step 4.

As is described in Step 2, the moment function $\mu_P(.)$ used in constructing the test statistic is indexed by $\ell = (x_1, x_2, q)$. A graphic illustration of different $\ell$ combinations and the resulting $[x_1, x_1 + \delta_x/q]$ and $[x_2, x_2 + \delta_x/q]$ intervals used in defining the moment functions are given in the bottom right graph of Figure 1. The choice of $Q$ determines the smallest cubes used in testing as well as the total number of moment functions involved,

which increases sharply with $Q$. When $X$ is single-dimensional, the count of unique $\ell$ combinations is equal to $(Q-1)Q(2Q-1)/12 + (Q-1)Q/4$. Andrews and Shi (2013, 2014) propose to choose $Q$ such that the expected sample size in the smallest cubes is around 20. We follow their advice in general. In empirical applications, however, such a rule of thumb might result in a computationally infeasible $Q$ value, like in our first empirical application. In such cases, we recommend carrying out the test using some large but still computationally feasible $Q$ values and also double check the robustness of testing results with respect to the $Q$ choice.

If the monotonicity relationship of interest is of the other direction, i.e., researchers are interested in testing whether $CLATE(.)$ is monotonically non-increasing, then one could redefine the outcome variable as $-Y_i$, reestimate $CLATE(.)$ with the new outcome variable, and then apply the above step-by-step implementation procedure. The proposed benchmark test could also be extended to examine heterogeneity treatment effects with multidimensional controls. We will consider such extensions in Section 4.

### 3.5 Testing under Sharp RD Design

When the treatment variable is a deterministic function of the running variable and the model is of sharp RD design, the hypotheses of interest could be formalized as

$$H_{0,SRD} : CATE(x) \text{ is non-decreasing in } x \text{ on } \mathcal{X};$$
$$H_{1,SRD} : \ H_{0,SRD} \text{ does not hold.}$$

This set of null and alternative hypotheses could also be used to examine whether the first-stage take-up rate of a fuzzy RD model changes monotonically with the value of $X$ at the RD cut-off $c$, as the first-stage take-up decision of any fuzzy RD model follows a sharp RD design.

Let $p_P^{(\kappa)}$ be a special case of $\varrho_P^{(\kappa)}$ with $T = 1(Z \geq c)$. Since sharp RD is a special case of fuzzy RD with full compliance, we know by Lemma 3.2 that testing the null $H_{0,SRD}$ is equivalent to testing

$$H_{0,SRD}^T : \nu_P(\ell) \equiv \rho_P^{(2)}(\ell)p_P^{(1)}(\ell) - \rho_P^{(1)}(\ell)p_P^{(2)}(\ell) \leq 0, \quad \text{for all } \ell \in \mathcal{L},$$

as long as $CATE(x)$ is continuous and $f_{X|Z}(x|z)$ is continuous and uniformly bounded. Test statistic and decision rules could also be constructed with the procedure for the FRD benchmark.

More efficiently, however, one could estimate $p_P^{(\kappa)}(\ell) = E_P[g_{x_\kappa,q}(X)|Z=c]$ by non-parametric regression using data from both sides of the RD cut-off. Let $\hat{p}_n^{(\kappa)}(\ell)$ be the intercept solution of the following minimization problem:

$$\left(\hat{p}_n^{(\kappa)}(\ell), \hat{b}_{n,p}^{(\kappa)}(\ell)\right) = \arg\min_{a,b} \sum_{i=1}^{n} K\left(\frac{Z_i - c}{h}\right)\left[g_{x_\kappa,q}(X_i) - a - b(Z_i - c)\right]^2.$$

For $l = 0, 1, 2$, define $S_{n,l} = \sum_{i=1}^{n} K(\frac{Z_i - c}{h})(Z_i - c)^l$ and $\mathsf{w}_{ni} = \frac{K(\frac{Z_i-c}{h})[S_{n,2} - S_{n,1}(Z_i - c)]}{S_{0,2}S_{n,2} - S_{n,1}^2}$. It then follows straightforwardly that $\hat{p}_n^{(\kappa)}(\ell) = \sum_{i=1}^{n} \mathsf{w}_{ni} \cdot p^{(\kappa)}(X_i, \ell)$ for both $\kappa = 1, 2$.

Let $\hat{\nu}_n(.) = \hat{\rho}_n^{(1)}(.)\hat{p}_n^{(2)}(.) - \hat{\rho}_n^{(2)}(.)\hat{p}_n^{(1)}(.)$ be the estimator of $\nu_P(.)$. Then the influence function representation of $\hat{\nu}_n(.)$ could be formulated as

$$\sqrt{nh}(\hat{\nu}_n(.) - \nu_P(.)) = \sum_{i=1}^{n} \phi_{\nu,ni}(.) + o_p(1),$$

where $\phi_{\nu,ni}(.)$ is defined similar to $\phi_{\mu,ni}(.)$ in Section 3.2 except that $\varrho_P^{(\kappa)}(.)$ and $\phi_{\varrho,ni}^{(\kappa)}(.)$ in $\phi_{\mu,ni}(.)$ are replaced by $p_P^{(\kappa)}(.)$ and $\phi_{p,ni}^{(\kappa)}(.) = \sqrt{nh}\left(\mathsf{w}_{ni}\left(p^{(\kappa)}(X_i, .) - p_P^{(\kappa)}(.)\right)\right)$ for $\kappa = 1, 2$. Let $\hat{\phi}_{\nu,ni}(.)$ being the estimated influence function and $\hat{\sigma}_{\nu,n}^2(.) = \sum_{i=1}^{n} \hat{\phi}_{\nu,ni}^2(.)$ be the variance estimator of $\hat{\nu}_n(.)$. Let $\hat{\sigma}_{\nu,\epsilon}^2(.) = \max\left\{\hat{\sigma}_{\nu,n}^2(.), \epsilon \cdot \hat{\sigma}_{\nu,n}^2(\ell_0)\right\}$. We can define the test statistic for the sharp RD case as

$$\widehat{T}_{n,SRD} = \sup_{\ell \in \mathcal{L}} \sqrt{nh} \frac{\hat{\nu}_n(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)}.$$

The LFC and GMS simulated critical values are defined as

$$\hat{c}_{n,SRD}^{\eta,LFC}(\alpha) = \sup\left\{q \middle| P^u\left(\sup_{\ell \in \mathcal{L}} \frac{\widehat{\Phi}_{\nu,n}^u(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)} \leq q\right) \leq 1 - \alpha + \eta\right\} + \eta, \text{ and}$$

$$\hat{c}_{n,SRD}^{\eta,GMS}(\alpha) = \sup\left\{q \middle| P^u\left(\sup_{\ell \in \mathcal{L}} \left(\frac{\widehat{\Phi}_{\nu,n}^u(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)} + \hat{\psi}_\nu(\ell)\right) \leq q\right) \leq 1 - \alpha + \eta\right\} + \eta$$

respectively, with $\widehat{\Phi}_{\nu,n}^u(\ell) = \sum_{i=1}^{n} U_i \cdot \hat{\phi}_{\nu,ni}(\ell)$ being the process simulating the limiting process of $\sqrt{nh}(\hat{\nu}_n(\ell) - \nu_P(\ell))$ and $\hat{\psi}_\nu(\ell) = -B_n \cdot 1\left(\sqrt{nh}\frac{\hat{\nu}_n(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)} < -a_n\right)$ be the GMS recentering term. If we reject the null hypothesis $H_{0,SRD}$ when $\widehat{T}_{n,SRD} > \hat{c}_{n,SRD}^{\eta,LFC}(\alpha)$ or when $\widehat{T}_{n,SRD} > \hat{c}_{n,SRD}^{\eta,GMS}(\alpha)$, the resulting tests are consistent and have uniform size control in the limit. The asymptotic properties are similar to those of the fuzzy RD test. We omit the details for brevity.

# 4 Extensions

The benchmark test considers the simplest form of CLATE(.) or CATE(.) with a single-dimensional continuous conditioning variable. In some empirical settings, it may be of interest to examine whether an RD treatment effect changes monotonically with some multidimensional random variable, while conditioning on values of a separate multidimensional control. In this section, we extend the benchmark test to consider this type of general treatment effect monotonicity relationship. We also discuss extensions to include discrete covariates in the testing procedure.

## 4.1 Multi-dimensional Effect Heterogeneity

Let $X$ be a $d_x$-dimensional random variable and $S$ be a $d_s$-dimensional random variable. $X \in \mathcal{X} = [0,1]^{d_x}$, $S \in \mathcal{S} = [0,1]^{d_s}$ and they include vectors of random variables that are mutually exclusive. Formally, define the null and alternative hypotheses of the general test as

$$\begin{aligned} &H'_{0,FRD} : CLATE(x,s) \text{ is non-decreasing in } x \text{ on } \mathcal{X} \text{ for all } s \in \mathcal{S}; \\ &H'_{1,FRD} : \ H'_{0,FRD} \text{ does not hold.} \end{aligned} \tag{4.1}$$

When $d_x = 1$ and $d_s = 0$, the general statements reduce to the hypotheses $H_{0,FRD}$ and $H_{1,FRD}$ in Section 3.1. When $d_x \geq 2$, we define the monotonic non-decreasing (or weakly monotonic increasing) relationship as $x_1 \geq x_2$ iff $x_{1j} \geq x_{2j}$ for all $j = 1, \ldots, d_x$. We also denote $x_1 > x_2$ iff $x_{1j} \geq x_{2j}$ for all $j = 1, \ldots, d_x$ and $x_{1k} > x_{2k}$ for at least one $k \in \{1, \ldots, d_x\}$, and denote $x_1 \gg x_2$ iff $x_{1j} > x_{2j}$ for all $j = 1, \ldots, d_x$.

Note that the statement in $H'_{0,FRD}$ is stronger when the control variable $S$ includes more elements. If $H'_{0,FRD}$ holds with some random variable $S$, then $H_{0,FRD}$ unconditional on $S$ must hold. But if $H'_{0,FRD}$ does not hold, $H_{0,FRD}$ can still hold true if $H'_{0,FRD}$ is violated in such a way that the violation could be canceled out when averaged over $S$. In the next, we first consider the case where both $X$ and $S$ are continuous. Discrete covariates will be discussed in the next section.

For any $x_1, x_2 \in \mathcal{X}$, $s \in \mathcal{S}$, and $q \in \{2, 3, \cdots\}$, define

$$C_{x_1,q} \equiv \prod_{j=1}^{d_x} [x_{1j}, x_{1j} + 1/q], \quad C_{x_2,q} \equiv \prod_{j=1}^{d_x} [x_{2j}, x_{2j} + 1/q],$$
$$C_{s,q} \equiv \prod_{j=1}^{d_s} [s_j, s_j + 1/(q-1)],$$

where $x_{1j}$, $x_{2j}$, and $s_j$ are the $j$-th dimension of $x_1$, $x_2$ and $s$. Let $g_{x_\kappa,q}(\cdot) = 1(\cdot \in C_{x_\kappa,q})$ for $\kappa = 1, 2$, $g_{s,q}(\cdot) = 1(\cdot \in C_{s,q})$, and $\ell = (x_1, x_2, s, q) \in \mathcal{X}^2 \times \mathcal{S} \times \{2, 3, \cdots\}$. Similar to Lemma 3.2, we have the following Lemma that transforms the null of general treatmnet effect monotonicity in $H'_{0,FRD}$ to a collection of moment inequalities without loss of information.

**Lemma 4.1** *Replace the random variable $X$ in Assumptions 2.1 and 3.1 with the new conditioning set $(X, S)$. Assume that the new conditions hold. Then the null hypothesis $H'_{0,FRD}$ is equivalent to*

$$H'^T_{0,FRD} : \mu_P(\ell) \equiv \rho_P^{(2)}(\ell)\varrho_P^{(1)}(\ell) - \rho_P^{(1)}(\ell)\varrho_P^{(2)}(\ell) \leq 0, \quad \text{for all } \ell \in \mathcal{L}, \tag{4.2}$$

*where*

$$\mathcal{L} = \Big\{(x_1, x_2, s, q) : \ (x_1, x_2) \in \{0, 1/q, 2/q, \cdots, (q-1)/q\}^{2d_x}, x_1 > x_2,$$
$$s \in \{0, 1/(q-1), 2/(q-1), \cdots, (q-2)/(q-1)\}^{d_s}, \ \text{for } q = 2, 3 \cdots \Big\},$$

*and*

$$\begin{aligned}
\varrho_P^{(\kappa)}(\ell) &\equiv q_{P,+}^{(\kappa)}(\ell) - q_{P,-}^{(\kappa)}(\ell) \\
&\equiv \lim_{z \searrow c} E_P[g_{x_\kappa,q}(W) g_{s,q}(S) T | Z = z] - \lim_{z \nearrow c} E_P[g_{x_\kappa,q}(W) g_{s,q}(S) T | Z = z] \\
\rho_P^{(\kappa)}(\ell) &\equiv m_{P,+}^{(\kappa)}(\ell) - m_{P,-}^{(\kappa)}(\ell) \\
&\equiv \lim_{z \searrow c} E_P[g_{x_\kappa,q}(W) g_{s,q}(S) Y | Z = z] - \lim_{z \nearrow c} E_P[g_{x_\kappa,q}(W) g_{s,q}(S) Y | Z = z]
\end{aligned}$$

*for $\kappa = 1, 2$.*

Replacing $g_{x_\kappa,q}(X_i)$ in Section 3.2 with $g_{x_\kappa,q}(X_i) g_{s,q}(S_i)$, we can construct the extended test for $H'_{0,FRD}$ based on Lemma 4.1 in the same way as the benchmark test. The extended test would enjoy the same asymptotic properties, including uniform size

control and test consistency. Proofs of the asymptotic properties are similar to those provided in the Appendix for the benchmark test. They are given in an earlier working paper version of the paper and are available upon request.

### 4.2 Discrete variables

This section extends the proposed testing procedure to include discrete random variables with finite distinct values.

First we consider the same null hypothesis $H_{0,FRD}'$ as in (4.1) except that the variable $X$ is now a mixture of continuous and discrete random variables. Combine all discrete elements of $X$ into a scalar random variable $X_{disc}$ with support $\mathcal{X}_{disc} = \{\ddot{x}_1, \ddot{x}_2, \ldots, \ddot{x}_M\}$. Since $X_{disc}$ is a part of the hetergeneity variable $X$ that is conjectured to have a possibly monotonic impact on the RD effect of interest, its value has to be ordered (e.g., years of schooling). Assume without loss of generality that $\ddot{x}_1 < \ddot{x}_2 < \cdots < \ddot{x}_M$. Let $X_{cts} \in [0,1]^{d_x - 1}$ be the $d_x - 1$ dimensional continuous element of $X$. The null hypothesis of interest would be

$$H_{0,FRD}^{''} : CLATE(x,s) \text{ is non-decreasing in } x \text{ on } \mathcal{X} = \mathcal{X}_{disc} \times \mathcal{X}_{disc} \text{ for all } s \in \mathcal{S}.$$

Because the above null hypothesis is the same as $H_{0,FRD}^{'}$ except for the new definition of $X$, this test extension could be carried out using the same null transformation strategy as in Lemma 4.1 but with new definitions of $\mathcal{L}$, $g_{x_1,q}(X)$, and $g_{x_2,q}(X)$. Let

$$\begin{aligned}\mathcal{L} = \Big\{ (x_1, x_2, s, q) : & (x_1^c, x_2^c) \in \{0, 1/q, \cdots, (q-1)/q\}^{2(d_x-1)}, x_1^d = \ddot{x}_{d_1}, x_2^d = \ddot{x}_{d_2}, \\ & (d_1, d_2) \in \{1, \cdots, K\}^2, x_1 = [x_1^c \ x_1^d]', x_2 = [x_2^c \ x_2^d]', x_1 > x_2, \\ & s \in \{0, 1/(q-1), 2/(q-1), \cdots, (q-2)/(q-1)\}^{d_s}, \text{ for } q = 2, 3 \cdots \Big\},\end{aligned}$$

and $g_{x_k,q}(X) = 1\left(X_{cts} \in \prod_{j=1}^{d_x-1}\left[x_{kj}^c, x_{kj}^c + q^{-1}\right]\right) \cdot 1\left(X_{disc} \in [\ddot{x}_{d_k}, \ddot{x}_{d_k+\Delta}]\right)$, where $\Delta = \min(d_2 - d_1 - 1, K - d_2)$. The new definitions of $\mathcal{L}$, $g_{x_1,q}(X)$, and $g_{x_2,q}(X)$ accomondate both continuous and discrete elements of $X$. Test statistic construction and critical value calculation follow the same procedures as described in Section 3 for the benchmark test.

Next we consider the same null hypothesis $H_{0,FRD}'$ as in (4.1) except that the last dimension of the conditioning variable $S$ is discrete. Let $S \equiv [S_{cts} \ S_{disc}]$, where $S_{cts} \in$

$[0,1]^{d_s-1}$ is continuous and $S_{disc} \in \mathcal{S}_{disc} = \{\ddot{s}_1, \ldots, \ddot{s}_M\}$ is discrete. Since $S$ is only a control variable in the hypothesis, $S_{disc}$ can be either ordered or unordered (e.g., race). Specifically, the null hypothesis of interest would be

$$H_{0,FRD}''' : CLATE(x,s) \text{ is non-decreasing in } x \text{ on } \mathcal{X} \text{ for all } s \in \mathcal{S} = [0,1]^{d_s-1} \times \mathcal{S}_{disc}.$$

Let $\mathfrak{S}$ be a collection of sets that include all $M$ distinct values of $\mathcal{S}_{disc}$ as well as $\mathcal{S}_{disc}$ itself. $\mathfrak{S}$ may also include other nonempty subsets of $\mathcal{S}_{disc}$. Suppose $\mathfrak{S}$ has $K$ distinct set elements denoted by $\{\mathfrak{S}_k\}_{p=1}^K$. The null hypothesis $H_{0,FRD}'''$ is equivalent to

$$H_{0,FRD-T}''' : \ \rho_P^{(2)}(\ell,k)\varrho_P^{(1)}(\ell,k) - \rho_P^{(1)}(\ell,k)\varrho_P^{(2)}(\ell,k) \leq 0, \text{ for all } \ell \in \mathcal{L} \text{ and } k = 1, ..., K, \tag{4.3}$$

where

$$\mathcal{L} = \Big\{(x_1, x_2, s^c, q) : (x_1, x_2) \in \{0, 1/q, \cdots, (q-1)/q\}^{2d_x},$$
$$s^c \in \{0, 1/(q-1), 2/(q-1), \cdots, (q-2)/(q-1)\}^{d_s-1}, \text{ for } q = 2, 3 \cdots \Big\},$$

and for both $\kappa = 1, 2$

$$\varrho_P^{(\kappa)}(\ell,k) = \lim_{z \searrow c} E_P[g_{x_\kappa,q}(X) g_{s,q,k}(S) T | Z = z] - \lim_{z \nearrow c} E_P[g_{x_\kappa,q}(X) g_{s,q,k}(S) T | Z = z],$$
$$\rho_P^{(\kappa)}(\ell,k) = \lim_{z \searrow c} E_P[g_{x_\kappa,q}(X) g_{s,q,k}(S) Y | Z = z] - \lim_{z \nearrow c} E_P[g_{x_\kappa,q}(X) g_{s,q,k}(S) Y | Z = z],$$

with $g_{x_\kappa,q}(X)$ as defined above and $g_{s,q,k}(S) = 1\left(S_{cts} \in \prod_{j=1}^{d_s-1}\left[s_j^c, s_j^c + (q-1)^{-1}\right]\right) \cdot 1\left(S_{disc} \in \mathfrak{S}_k\right)$. Given the moment inequalities in the transformed null hypothesis, test statistic construction and critical value calculation follow the same procedures as described in Section 3. The collection of subset $\mathfrak{S}$ could be the sigma-algebra of $\mathcal{S}_{disc}$ if its cardinal count, or $M$, is relatively small. Otherwise, $\mathfrak{S} = \{\{\ddot{s}_1\}, \ldots, \{\ddot{s}_M\}, \mathcal{S}_{disc}\}$ would be computationally feasible and yet enough to maintain the consistency property of the monotonicity test.

The general case where both $X$ and $S$ include discrete elements could be tested combining the above two extensions. The details are omitted for brevity.

# 5 Simulations

This section examines the small sample performance of the proposed tests using Monte Carlo simulations. For all data generating processes (DGPs), the running variable $Z$,

the additional control $X$, and the error term $u$ in the outcome equation are generated as follows:

$$Z \sim 2Beta(2,2) - 1;\ X \sim U[0,1];\ u \sim N(0,1).$$

The outcome $Y$ and the treatment decision $T$ are DGP specific and are either estimated from empirical data or modified from the data-driven DGPs to demonstrate specific properties of the proposed tests. DGPs 5 and 6 also include a control variable $S$ that takes values 0 and 1 with equal probabilities.

First, we consider the sharp RD design in DGPs 1-4. The DGPs are plotted in Figure 1. DGPs 1-2 are estimated from empirical data with detailed DGP generating procedures described in the footnote of Figure 1. DGP 3 and DGP 4 are altered from DGP 2 to show problems of the interaction term method that is popular in the applied literature. DGP 3 also demonstrates the potential power gain from using the GMS critical value. For each DGP, three different sample sizes, $n = 2,000$, $n = 4,000$ and $n = 8,000$, are used.

*DGP 1: Sharp RD, Constant Effect, Homogeneous Model*

$$Y = \begin{cases} -0.373 + 0.545Z - 0.056Z^2 + 0.1u & \text{if } Z \geq 0 \\ -0.531 + 0.556Z + 0.192Z^2 + 0.1u & \text{if } Z < 0 \end{cases}$$

*DGP 2: Sharp RD, Monotonically Increasing Treatment Effect*

$$Y = \begin{cases} -0.921 + 0.833X + 0.584Z - 0.054Z^2 + 0.1u & \text{if } Z \geq 0 \\ -0.705 + 0.264X + 0.580Z + 0.191Z^2 + 0.1u & \text{if } Z < 0 \end{cases}.$$

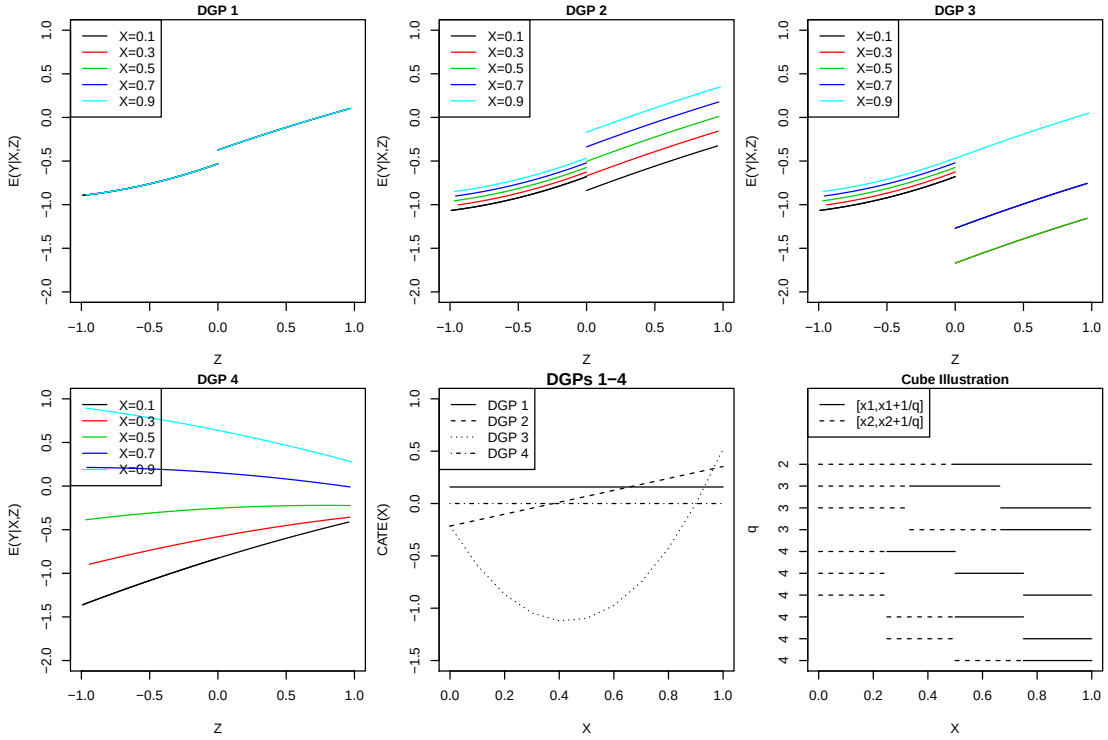
*DGP 3: Sharp RD, Inverse-U Shape Treatment Effect*

$$Y = \begin{cases} -0.921 - 4X + 0.584Z - 0.054Z^2 + 5X^2 + 0.1u & \text{if } Z \geq 0 \\ -0.705 + 0.264X + 0.580Z + 0.191Z^2 + 0.1u & \text{if } Z < 0 \end{cases}.$$

*DGP 4: Sharp RD, Constant Effect, Heterogeneous Model*

$$Y = \begin{cases} -0.921 + 0.833X + 0.584Z - XZ - 0.054Z^2 + X^2 + 0.1u & \text{if } Z \geq 0 \\ -0.921 + 0.833X + 0.584Z - XZ - 0.054Z^2 + X^2 + 0.1u & \text{if } Z < 0 \end{cases}.$$

Figure 1: DGP 1-3




Note: The DGPs are estimated from the data of the empirical section. To obtain the models, we first rescale the running variable, i.e., the standardized transition score, in the data set to $[-1, 1]$ to match the support of the generated $X$ variable. Then for the outcome equation in DGP 1, we regress the score in the Baccalaureate exam on the running variable and its second-order polynomial term separately for the subsample to the left and the right of the cutoff value. To get the outcome equation in DGP 2, we add the additional control of interest, i.e., average peer admission score into the regression. DGP 3 and DGP 4 are altered from DGP 2.

Two null hypotheses are considered for each DGP, testing whether the $CATE(x)$ is monotonically non-decreasing in $x$, and whether $CATE(x)$ is monotonically non-increasing in $x$, respectively. For each DGP, 1,000 simulation repetitions are carried out, and in each monotonicity test, critical values are calculated using 1,000 bootstrap simulations. All tests carried out in this section use the triangular kernel. Bandwidth selection follows the rule-of-thumb recommendation in Section 3.4 with $k = 4.25, 4.5, 4.75$.

The moment function $\mu_P(\ell)$ configuration also follows the discussion in Section 3.4. A graphic illustration of different $\ell$ combinations and the resulting $[x_1, x_1 + \delta_x/q]$ and $[x_2, x_2 + \delta_x/q]$ intervals used in testing are given in the bottom right graph of Figure 1, for $q = 2, 3, 4$. In Monte Carlo simulations, moment functions are configured with $Q = 10$ and 15. When $Q = 10$, a total of 165 unique moment functions are used to construct the

test statistic. The number is 560 when $Q = 15$. With those $Q$ values, the smallest effective sample size in kernel estimations averages around 32 when $Q = 10$ and $n = 2,000$, and around 21 when $Q = 15$ and $n = 2,000$.

Table 1 summarizes the rejection proportions of the proposed tests for the sharp RD case, with $p_P^{(\kappa)}(\ell)$ estimated using the full sample method as is discussed in Section 3.5. We see from the table that, regardless of whether the LFC or the GMS critical value is used, the proposed monotonicity test controls size well in small samples and has power approaching one as the sample size increases. The GMS method brings the size of the proposed test closer to the nominal level (5%) and increases power when the null hypothesis of interest is violated. The power gain of using the GMS critical value is larger with DGP 3 than with DGP 2 (for the null of a non-increasing relationship) because with the inverse U-shaped CATE in DGP 3 the GMS method can get rid of the influence of about half of the simulated control functions in critical value calculation when testing either direction of monotonicity. Test results are also qualitatively similar for both choices of $Q$ values, although power is typically higher with smaller $Q$ values for this set of DGPs. This is because none of the CATE functions in the DGPs have an overly oscillating pattern that requires large $Q$ to obtain consistency. Under our DGPs, $Q = 10$ is more than enough to capture the violation, if any. Larger $Q$ values would then result in lower finite sample power, as more moment functions need to be used in testing. In empirical applications, it is not necessary for the tests to have higher p-values when $Q$ is larger, as the conditional mean functions in the real data could be noisier.

Table 1: Small Sample Performance of the Benchmark Test Under Sharp RD

| | LFC Critical Value | | | | | | GMS Critical Value | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_0$ : | Non-decreasing CATE | | | Non-increasing CATE | | | Non-decreasing CATE | | | Non-increasing CATE | | |
| $k$ | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 |
| | | | | | | Panel A: $Q = 10$ | | | | | | |
| | DGP 1: Sharp RD, Constant Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.012 | 0.018 | 0.015 | 0.022 | 0.021 | 0.024 | 0.013 | 0.019 | 0.017 | 0.024 | 0.027 | 0.028 |
| n=4000 | 0.028 | 0.035 | 0.038 | 0.025 | 0.025 | 0.025 | 0.030 | 0.038 | 0.042 | 0.027 | 0.028 | 0.027 |
| n=8000 | 0.032 | 0.038 | 0.042 | 0.037 | 0.032 | 0.035 | 0.037 | 0.042 | 0.044 | 0.038 | 0.035 | 0.039 |
| | DGP 2: Sharp RD, Monotonically Increasing Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.003 | 0.002 | 0.003 | 0.136 | 0.162 | 0.182 | 0.004 | 0.004 | 0.005 | 0.143 | 0.167 | 0.189 |
| n=4000 | 0.007 | 0.005 | 0.008 | 0.375 | 0.427 | 0.475 | 0.009 | 0.008 | 0.010 | 0.381 | 0.431 | 0.480 |
| n=8000 | 0.003 | 0.002 | 0.001 | 0.705 | 0.776 | 0.830 | 0.004 | 0.003 | 0.002 | 0.709 | 0.776 | 0.831 |
| | DGP 3: Sharp RD, Inverse U-Shaped Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.038 | 0.044 | 0.046 | 0.431 | 0.505 | 0.579 | 0.054 | 0.059 | 0.063 | 0.455 | 0.525 | 0.593 |
| n=4000 | 0.085 | 0.110 | 0.126 | 0.902 | 0.934 | 0.960 | 0.127 | 0.153 | 0.168 | 0.911 | 0.938 | 0.960 |
| n=8000 | 0.272 | 0.301 | 0.337 | 1.000 | 1.000 | 1.000 | 0.326 | 0.368 | 0.406 | 1.000 | 1.000 | 1.000 |
| | DGP 4: Sharp RD, Constant Treatment Effect With Heterogeneous Model | | | | | | | | | | | |
| n=2000 | 0.030 | 0.030 | 0.036 | 0.026 | 0.031 | 0.035 | 0.035 | 0.036 | 0.037 | 0.030 | 0.035 | 0.038 |
| n=4000 | 0.034 | 0.033 | 0.037 | 0.035 | 0.036 | 0.037 | 0.038 | 0.037 | 0.041 | 0.038 | 0.047 | 0.046 |
| n=8000 | 0.041 | 0.051 | 0.052 | 0.040 | 0.039 | 0.043 | 0.043 | 0.052 | 0.056 | 0.045 | 0.043 | 0.047 |
| | | | | | | Panel B: $Q = 15$ | | | | | | |
| | DGP 1: Sharp RD, Constant Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.008 | 0.008 | 0.010 | 0.011 | 0.014 | 0.016 | 0.010 | 0.008 | 0.011 | 0.012 | 0.016 | 0.017 |
| n=4000 | 0.020 | 0.028 | 0.027 | 0.017 | 0.019 | 0.020 | 0.026 | 0.032 | 0.034 | 0.021 | 0.020 | 0.024 |
| n=8000 | 0.028 | 0.031 | 0.036 | 0.032 | 0.032 | 0.032 | 0.032 | 0.035 | 0.037 | 0.036 | 0.033 | 0.034 |
| | DGP 2: Sharp RD, Monotonically Increasing Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.001 | 0.002 | 0.002 | 0.102 | 0.115 | 0.132 | 0.001 | 0.002 | 0.003 | 0.103 | 0.121 | 0.136 |
| n=4000 | 0.004 | 0.005 | 0.006 | 0.300 | 0.354 | 0.401 | 0.007 | 0.007 | 0.009 | 0.304 | 0.361 | 0.406 |
| n=8000 | 0.002 | 0.002 | 0.002 | 0.649 | 0.733 | 0.784 | 0.004 | 0.003 | 0.004 | 0.651 | 0.735 | 0.786 |
| | DGP 3: Sharp RD, Inverse U-Shaped Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.026 | 0.033 | 0.034 | 0.327 | 0.400 | 0.484 | 0.034 | 0.040 | 0.041 | 0.350 | 0.428 | 0.516 |
| n=4000 | 0.073 | 0.082 | 0.104 | 0.847 | 0.904 | 0.935 | 0.090 | 0.117 | 0.135 | 0.859 | 0.909 | 0.939 |
| n=8000 | 0.235 | 0.274 | 0.310 | 0.999 | 1.000 | 1.000 | 0.281 | 0.322 | 0.354 | 0.999 | 1.000 | 1.000 |
| | DGP 4: Sharp RD, Constant Treatment Effect With Heterogeneous Model | | | | | | | | | | | |
| n=2000 | 0.016 | 0.014 | 0.018 | 0.010 | 0.011 | 0.012 | 0.019 | 0.019 | 0.019 | 0.010 | 0.013 | 0.016 |
| n=4000 | 0.015 | 0.020 | 0.025 | 0.020 | 0.019 | 0.019 | 0.020 | 0.022 | 0.028 | 0.022 | 0.023 | 0.024 |
| n=8000 | 0.034 | 0.038 | 0.041 | 0.033 | 0.039 | 0.033 | 0.039 | 0.045 | 0.045 | 0.040 | 0.044 | 0.038 |

Note: Reported are rejection proportions among 1,000 simulations, where all tests are carried out using the 5% significance level. For each test, the simulated critical value is calculated with 1,000 bootstrap repetitions.

Table 2 reports testing results from the interaction term method that is popularly adopted in empirical RD studies for heterogeneity analysis. The method involves adding an interaction term between the heterogeneity variable $X$ and the dummy variable $1(Z \geq c)$ into the classic RD regression, which, for the sharp RD models considered here, is weighted OLS of $Y$ on $1(Z \geq c)$, $Z$, and $1(Z \geq c) \cdot Z$. OLS weights are determined by $K((Z-c)/h)$. The interaction term method is parametric because the nonparametric estimation of $CATE(x)$ involves conditioning on both $X = x$ and $Z = c$, while the interaction term method only uses weights calculated based on the distance of $Z$ from the cutoff $c$. We see from the table that monotonicity tests based on the interaction term method (i.e. one-sided t-tests of the slope on the interaction term) over-rejects badly under DGP 4. Also, under DGP 3, when the $CATE(x)$ function first decreases then increases with the value of $x$, the interaction term method often suggests a false monotonic functional form of the $CATE$ function.

Table 2: Performance of the Interaction Term Method

| $H_0$ : | Non-decreasing CATE | | | Non-increasing CATE | | | Non-decreasing CATE | | | Non-increasing CATE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 |
| | DGP 1 | | | | | | DGP 3 | | | | | |
| n=2000 | 0.069 | 0.069 | 0.071 | 0.068 | 0.065 | 0.063 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 |
| n=4000 | 0.089 | 0.089 | 0.088 | 0.077 | 0.072 | 0.070 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 |
| n=8000 | 0.072 | 0.074 | 0.076 | 0.069 | 0.063 | 0.068 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 |
| | DGP 2 | | | | | | DGP 4 | | | | | |
| n=2000 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 | 0.934 | 0.971 | 0.987 | 0.000 | 0.000 | 0.000 |
| n=4000 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 | 0.978 | 0.989 | 0.994 | 0.000 | 0.000 | 0.000 |
| n=8000 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 | 0.996 | 1.000 | 1.000 | 0.000 | 0.000 | 0.000 |

Note: Reported are rejection proportions among 1,000 simulations, where all tests are carried out using the 5% significance level. For each test, the simulated critical value is calculated with 1,000 bootstrap repetitions.

Table 3 repeats the simulation experiments in panel A of Table 1 but with undersmoothed Imbens and Kalyanaraman (2012) (IK) bandwidths rather than undersmoothed CCT bandwidths suggested in Section 3.5. Results are qualitatively similar to those reported in Table 1, except that the test seems to have a small tendency of over-rejection (DGP1, $n = 8,000$) when alternative undersmoothed IK bandwidths are used. This finding is in line with the simulations results reported in CCT. Therefore, in

the rest of the paper, we will focus on the under-smoothed CCT bandwidth suggested in Section 3.4.

Table 3: Small Sample Performance of the Benchmark Test Under Sharp RD, Alternative Bandwidth Definition

| | LFC Critical Value | | | | | | GMS Critical Value | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_0$ : | Non-decreasing CATE | | | Non-increasing CATE | | | Non-decreasing CATE | | | Non-increasing CATE | | |
| $k$ | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 |
| | DGP 1: Sharp RD, Constant Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.021 | 0.024 | 0.025 | 0.026 | 0.024 | 0.027 | 0.022 | 0.025 | 0.029 | 0.027 | 0.027 | 0.029 |
| n=4000 | 0.021 | 0.026 | 0.032 | 0.032 | 0.033 | 0.030 | 0.027 | 0.028 | 0.033 | 0.036 | 0.037 | 0.034 |
| n=8000 | 0.047 | 0.051 | 0.053 | 0.045 | 0.050 | 0.048 | 0.049 | 0.053 | 0.057 | 0.050 | 0.053 | 0.053 |
| | DGP 2: Sharp RD, Monotonically Increasing Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.005 | 0.007 | 0.006 | 0.217 | 0.255 | 0.287 | 0.010 | 0.008 | 0.011 | 0.222 | 0.259 | 0.290 |
| n=4000 | 0.005 | 0.007 | 0.004 | 0.531 | 0.583 | 0.625 | 0.012 | 0.009 | 0.009 | 0.535 | 0.586 | 0.625 |
| n=8000 | 0.006 | 0.006 | 0.004 | 0.866 | 0.907 | 0.936 | 0.009 | 0.007 | 0.008 | 0.867 | 0.907 | 0.936 |
| | DGP 3: Sharp RD, Inverse U-Shaped Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.063 | 0.064 | 0.071 | 0.696 | 0.763 | 0.812 | 0.082 | 0.089 | 0.103 | 0.710 | 0.772 | 0.830 |
| n=4000 | 0.163 | 0.181 | 0.216 | 0.970 | 0.981 | 0.988 | 0.210 | 0.233 | 0.271 | 0.971 | 0.981 | 0.989 |
| n=8000 | 0.388 | 0.427 | 0.489 | 0.999 | 0.999 | 0.999 | 0.466 | 0.511 | 0.563 | 0.999 | 0.999 | 0.999 |
| | DGP 4: Sharp RD, Constant Treatment Effect With Heterogeneous Model | | | | | | | | | | | |
| n=2000 | 0.022 | 0.024 | 0.027 | 0.034 | 0.034 | 0.036 | 0.031 | 0.033 | 0.036 | 0.041 | 0.039 | 0.042 |
| n=4000 | 0.028 | 0.034 | 0.036 | 0.035 | 0.042 | 0.045 | 0.037 | 0.036 | 0.043 | 0.038 | 0.050 | 0.052 |
| n=8000 | 0.039 | 0.036 | 0.032 | 0.046 | 0.048 | 0.051 | 0.043 | 0.039 | 0.038 | 0.053 | 0.052 | 0.052 |

Note: Reported are rejection proportions among 1,000 simulations, where all tests are carried out using the 5% significance level. For each test, the simulated critical value is calculated with 1,000 bootstrap repetitions.

DGPs 5 and 6 illustrate the small sample performance of the proposed monotonicity tests when an additional discrete control variable $S$ is added to the model. $S$ takes a value of either 0 or 1, each with probability 0.5. DGP 5 is the same as DGP 1 regardless of the value of $S$. In other words, $S$ is irrelevant in the outcome equation. DGP 6 is the same as DGP 2 except that the slope of $X$ in the outcome equation becomes zero when $S = 0$. So the RD model in DGP 6 has a monotonically increasing treatment effect when $S = 1$ but a constant treatment effect when $S = 0$. Table 4 reports the rejection rates of the proposed test in Section 4.2 with $Q = 10$. With the additional control variable $S$, the smallest effective sample size of kernel estimation averages around 16 when $n = 2,000$. We see that the proposed test controls size well when the null hypothesis holds in DGP

5 and has decent power in DGP 6 for the null of non-increasing CATE even though the null is only violated at half of the $S$ values.

Table 4: Small Sample Performance of the Proposed Test with Discrete Control, $Q = 10$

| | LFC Critical Value | | | | | | GMS Critical Value | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_0$ : | Non-decreasing CATE | | | Non-increasing CATE | | | Non-decreasing CATE | | | Non-increasing CATE | | |
| $k$ | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 |
| | DGP 5: Sharp RD, Constant Effect Regardless of $S$ | | | | | | | | | | | |
| n=2000 | 0.011 | 0.015 | 0.016 | 0.008 | 0.012 | 0.016 | 0.011 | 0.015 | 0.016 | 0.011 | 0.016 | 0.017 |
| n=4000 | 0.019 | 0.025 | 0.026 | 0.015 | 0.018 | 0.025 | 0.021 | 0.026 | 0.031 | 0.016 | 0.021 | 0.026 |
| n=8000 | 0.038 | 0.042 | 0.043 | 0.029 | 0.030 | 0.035 | 0.039 | 0.046 | 0.044 | 0.031 | 0.033 | 0.037 |
| | DGP 6: Sharp RD, Constant Effect When $S = 0$, Monotonically Increasing Effect When $S = 1$ | | | | | | | | | | | |
| n=2000 | 0.007 | 0.009 | 0.011 | 0.080 | 0.090 | 0.103 | 0.008 | 0.010 | 0.012 | 0.084 | 0.094 | 0.109 |
| n=4000 | 0.012 | 0.016 | 0.017 | 0.220 | 0.265 | 0.306 | 0.017 | 0.017 | 0.018 | 0.226 | 0.275 | 0.311 |
| n=8000 | 0.015 | 0.021 | 0.015 | 0.530 | 0.606 | 0.672 | 0.017 | 0.021 | 0.024 | 0.537 | 0.612 | 0.675 |

Note: Reported are rejection proportions among 1,000 simulations, where all tests are carried out using the 5% significance level. For each test, the simulated critical value is calculated with 1,000 bootstrap repetitions.

The last three DGPs illustrate the small sample performance of the proposed monotonicity test under the fuzzy RD design. The outcome equations are the same as those in DGPs 1-3, respectively. The selection equation is modeled by $T = 1(0.331 + 0.277Z + 0.049Z^2 + u > 0)$ if $Z \geq 0$ and $T = 0$ otherwise, which is estimated from the data using probit regression. Table 5 summarizes the rejection proportions of the tests. The table reports small sample performance similar to those reported in Table 1, although the tests for the new DGPs generally have lower power due to the extra noise in the first stage.

Table 5: Small Sample Performance of Proposed Monotonicity Tests Under Fuzzy RD

| | LFC Critical Value | | | | | | GMS Critical Value | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_0$: | Non-decreasing CLATE | | | Non-increasing CLATE | | | Non-decreasing CLATE | | | Non-increasing CLATE | | |
| k | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 |
| | | | | | | $Q = 10$ | | | | | | |
| | DGP 7: Fuzzy RD, Constant Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 | 0.003 | 0.000 | 0.000 | 0.001 | 0.001 | 0.001 | 0.003 |
| n=4000 | 0.003 | 0.004 | 0.005 | 0.007 | 0.007 | 0.008 | 0.003 | 0.004 | 0.005 | 0.007 | 0.007 | 0.008 |
| n=8000 | 0.011 | 0.016 | 0.014 | 0.013 | 0.012 | 0.012 | 0.012 | 0.018 | 0.019 | 0.014 | 0.012 | 0.014 |
| | DGP 2: Fuzzy RD, Monotonically Increasing Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.000 | 0.000 | 0.000 | 0.054 | 0.067 | 0.079 | 0.000 | 0.000 | 0.000 | 0.056 | 0.071 | 0.079 |
| n=4000 | 0.000 | 0.000 | 0.000 | 0.192 | 0.237 | 0.281 | 0.001 | 0.000 | 0.000 | 0.200 | 0.240 | 0.284 |
| n=8000 | 0.000 | 0.001 | 0.001 | 0.531 | 0.616 | 0.683 | 0.001 | 0.001 | 0.001 | 0.534 | 0.617 | 0.685 |
| | DGP 3: Fuzzy RD, Inverse U-Shaped Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.012 | 0.021 | 0.025 | 0.183 | 0.244 | 0.307 | 0.022 | 0.032 | 0.035 | 0.198 | 0.257 | 0.326 |
| n=4000 | 0.062 | 0.079 | 0.092 | 0.749 | 0.840 | 0.896 | 0.086 | 0.110 | 0.136 | 0.765 | 0.851 | 0.905 |
| n=8000 | 0.291 | 0.359 | 0.419 | 0.996 | 0.999 | 1.000 | 0.392 | 0.460 | 0.531 | 0.997 | 0.999 | 1.000 |
| | | | | | | $Q = 15$ | | | | | | |
| | DGP 7: Fuzzy RD, Constant Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| n=4000 | 0.002 | 0.002 | 0.002 | 0.005 | 0.003 | 0.002 | 0.002 | 0.002 | 0.002 | 0.005 | 0.004 | 0.003 |
| n=8000 | 0.006 | 0.007 | 0.010 | 0.002 | 0.006 | 0.009 | 0.007 | 0.009 | 0.011 | 0.003 | 0.007 | 0.009 |
| | DGP 8: Fuzzy RD, Monotonically Increasing Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.000 | 0.000 | 0.000 | 0.030 | 0.041 | 0.052 | 0.000 | 0.000 | 0.000 | 0.032 | 0.044 | 0.053 |
| n=4000 | 0.000 | 0.000 | 0.000 | 0.135 | 0.169 | 0.200 | 0.000 | 0.000 | 0.000 | 0.138 | 0.172 | 0.204 |
| n=8000 | 0.000 | 0.000 | 0.000 | 0.460 | 0.530 | 0.607 | 0.000 | 0.000 | 0.001 | 0.464 | 0.532 | 0.613 |
| | DGP 9: Fuzzy RD, Inverse U-Shaped Treatment Effect | | | | | | | | | | | |
| n=2000 | 0.005 | 0.006 | 0.011 | 0.116 | 0.150 | 0.193 | 0.007 | 0.010 | 0.016 | 0.121 | 0.163 | 0.210 |
| n=4000 | 0.038 | 0.047 | 0.059 | 0.620 | 0.726 | 0.810 | 0.054 | 0.066 | 0.078 | 0.641 | 0.744 | 0.825 |
| n=8000 | 0.194 | 0.260 | 0.330 | 0.988 | 0.999 | 0.999 | 0.286 | 0.354 | 0.422 | 0.990 | 0.999 | 0.999 |

Note: Reported are rejection proportions among 1,000 simulations, where all tests are carried out using the 5% significance level. For each test, the simulated critical value is calculated with 1,000 bootstrap repetitions.

# 6 Empirical Examples

## 6.1 The Effect of Going to a Better High School

In this application, we revisit the regression discontinuity analysis in Pop-Eleches and Urquiola (2013) on the effect of going to a better high school in Romania. As is discussed

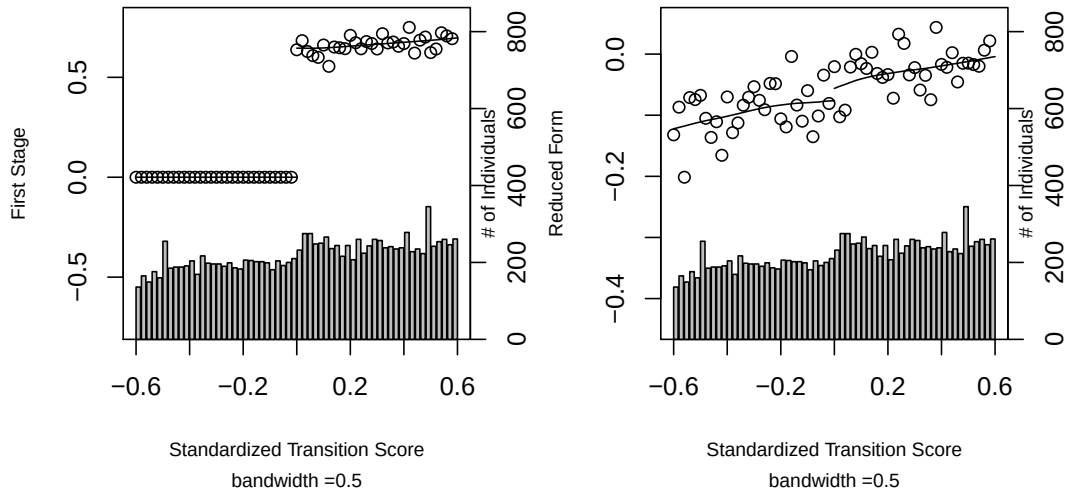
in Pop-Eleches and Urquiola (2013), elementary school students in Romania take a nationwide test in the last year of elementary school and apply to a list of high schools (and tracks). The admission decision is made based on a student's transition score, which is an average of the student's performance on the national test and his/her grade point average (GPA). A student is eligible for a high school if his/her transition score passes the school's cutoff. Using an administrative dataset, Pop-Eleches and Urquiola (2013) find that attending a better school on average improves a student's performance on the Baccalaureate exam but does not significantly affect his/her probability of taking the exam.

In this section, we focus on the outcome variable indicating whether a student takes the Baccalaureate exam ($Y$) and apply the proposed monotonicity test to examine whether the effect of attending a more selective high school ($T$) monotonically changes with the peer quality of the school ($X$) measured by the leave-one-out average transition score. The running variable ($Z$) in this application is a student's standardized transition score subtracting the individual school cutoff. As in Shen and Zhang (2016) and Hsu and Shen (2019), we focus on two-school towns because score cutoffs from different high schools within a town are often quite close to each other, and having more than one discontinuity point within the estimation window can introduce severe estimation bias.

Figure 2 provides a graphic summary of the data. In the left graph, the first-stage compliance rate is plotted against the histogram of the standardized transition score. No students go to the better high school with a transition score lower than the school-specific score cutoff. About 65% of the students with the transition score barely passing the cutoff choose to go to the more selective school. In the right graph, the conditional probability of taking the Baccalaureate exam (conditioning on student's transition score) jumps at the transition score cutoff. However, the jump is small in magnitude, and the data points of conditional exam-taking probabilities are noisy. Not surprisingly, Pop-Eleches and Urquiola (2013) report a statistically insignificant RD effect for this outcome variable.

Table 6 reports the results of the proposed nonparametric monotonicity tests. As is discussed in the simulation section, the test is carried out using the triangular kernel, an undersmoothed CCT bandwidth, and the moment transformation detailed in the implementation procedure in Section 3.4. We set $Q$ to 50 and 75 and find the empirical

Figure 2: Graphic Summary of Data




Note: Data are from Pop-Eleches and Urquiola (2013). Local linear regressions are carried out using the triangular kernel and fixed bandwidth 0.5.

results not very sensitive to the $Q$ choice. When $Q = 75$, the null transformation of our test involves a total of 70,300 different moment functions and the smallest effective sample size in kernel regressions used in our proposed testing procedure averages around 40.

First, we examine whether peer quality of the more selective school affects the enrollment decision of marginal students who barely pass the score cutoff of the more selective school. Our test rejects the null of a non-decreasing (or weakly increasing) first-stage effect with extremely close to zero p-values, and rejects the null of a non-increasing (or weakly decreasing) first-stage effect at the 10% significant value (based on p-values calculated from the GMS method). The testing results suggest a heterogeneous but not monotonic first-stage effect. It is also worth pointing out that the parametric interaction term method, in this case, estimates a positive and statistically significant slope on the interaction term. Therefore, the use of this parametric method could lead to a false conclusion that the first-stage take-up decision increases monotonically with the peer quality of the more selective high school.

Table 6: P-values of Monotonicity Tests

| | P-values of Proposed Tests | | | | | | Est. (St. Dev.) of the | | |
|---|---|---|---|---|---|---|---|---|---|
| | LFC | | | GMS | | | Interaction Term Method | | |
| k | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 |
| Panel A: | First-Stage Effect, Weakly Increasing | | | | | | Parametric Estimate (St. Dev.) | | |
| $Q = 50$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.107*** | 0.106*** | 0.107*** |
| $Q = 75$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | (0.021) | (0.019) | (0.018) |
| | First-Stage Effect, Weakly Decreasing | | | | | | | | |
| $Q = 50$ | 0.096 | 0.112 | 0.114 | 0.073 | 0.081 | 0.090 | - | - | - |
| $Q = 75$ | 0.096 | 0.112 | 0.114 | 0.073 | 0.081 | 0.090 | - | - | - |
| Panel B: | Reduced-Form Effect, Weakly Increasing | | | | | | Parametric Estimate (St. Dev.) | | |
| $Q = 50$ | 0.991 | 0.986 | 0.989 | 0.988 | 0.983 | 0.983 | 0.069*** | 0.067*** | 0.064*** |
| $Q = 75$ | 0.991 | 0.986 | 0.990 | 0.988 | 0.983 | 0.984 | (0.025) | (0.023) | (0.022) |
| | Reduced-Form Effect, Weakly Decreasing | | | | | | | | |
| $Q = 50$ | 0.017 | 0.040 | 0.031 | 0.017 | 0.040 | 0.031 | - | - | - |
| $Q = 75$ | 0.017 | 0.040 | 0.031 | 0.017 | 0.040 | 0.031 | - | - | - |
| Panel C | CLATE, Weakly Increasing | | | | | | Parametric Estimate (St. Dev.) | | |
| $Q = 50$ | 0.943 | 0.977 | 0.977 | 0.925 | 0.971 | 0.970 | 0.134*** | 0.123*** | 0.114*** |
| $Q = 75$ | 0.943 | 0.971 | 0.972 | 0.925 | 0.965 | 0.965 | (0.054) | (0.050) | (0.045) |
| | CLATE, Weakly Decreasing | | | | | | | | |
| $Q = 50$ | 0.036 | 0.035 | 0.025 | 0.036 | 0.035 | 0.025 | - | - | - |
| $Q = 75$ | 0.036 | 0.035 | 0.025 | 0.036 | 0.035 | 0.025 | - | - | - |

Notes: Data are from Pop-Eleches and Urquiola (2013). Nonparametric test statistics are calculated using the triangular kernel, the undersmoothed CCT bandwidth and null transformation defined in Section 3.4. The $Q$ choice is irrelevant for the parametric method. All simulated critical values are calculated with 1,000 bootstrap repetitions. The interaction term method is defined in Section 5.

Panel C of Table 6 examines the functional form of the conditional average treatment effect. Our proposed (fuzzy RD) monotonicity test rejects the null of a non-increasing effect with a 5% significance level, and the result is robust to the multiple bandwidths and $Q$ choices we worked with. Meanwhile, the proposed test also fails to reject the null of a non-decreasing effect with very high p-values despite the large sample size used in the analysis. Our testing results, therefore, suggest that the impact of attending a better high school on a marginal student's probability of taking the Baccalaureate exam increases with the peer quality of the more selective high school. Panel B of Table 6 reports testing results of the reduced-form RD effect as a robustness check. A similar monotonicity

relationship with peer quality was discovered. The parametric interaction term method also reports similar monotonic relationships in Panels B and C. The monotonic functional form consistently found in Panels B and C indicates that the insignificant mean effect found in Pop-Eleches and Urquiola (2013) and described earlier in Figure 2 might come from the cancelation of positive and negative treatment effects among different schools. Such an empirical finding is also in line with the results of Hsu and Shen (2019), who find that the effect on the Baccalaureate exam-taking rate is positive for some subpopulation of schools and negative for the others.

Pop-Eleches and Urquiola (2013) also examines the effect of going to a better high school on the scores of exam-takers on the Baccalaureate exam. We do not look at this outcome. Testing results reported in Panels B and C of Table 6 suggest non-random sample selection in taking the Baccalaureate exam. With additional assumptions, our monotonicity test could be extended to the sample selection set-up discussed in Dong (2019). However, such an extension is beyond the scope of this paper and is left for future research.

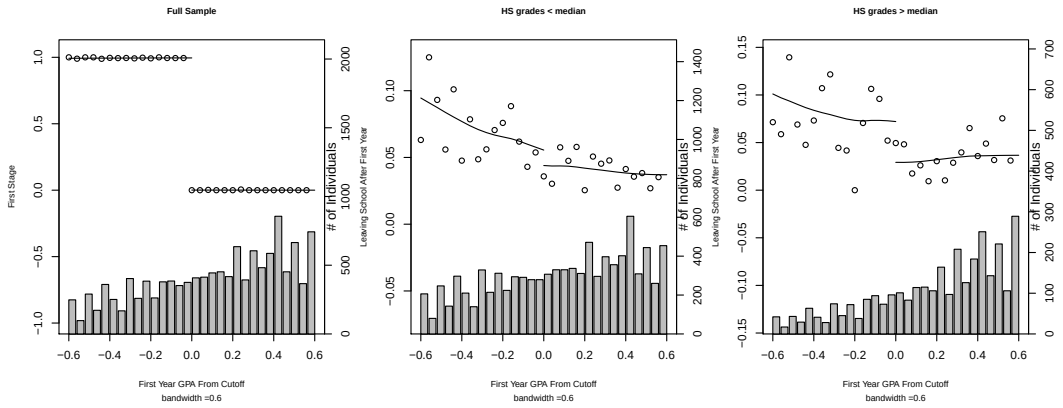
### 6.2 The Effect of Academic Probation in College

In this application, we revisit the regression discontinuity analysis in Lindo et al. (2010) on the effect of academic probation in college. Lindo et al. (2010) use data from a large Canadian university and find that being placed on academic probation at the end of the first year of college increases the probability of a marginal student leaving school, decreases his/her chance of graduation but improves the GPA of those who choose to stay. Lindo et al. (2010) also find the effects to be larger if the student's high school GPA is higher, or if he is male, or is a native English speaker.

In this section, we focus on the outcome ($Y$) of whether a student leaves school after the first year. The running variable ($Z$) is a student's first-year GPA subtracting the probation cutoff. The treatment variable ($T$) indicates whether a student receives academic probation at the end of the first year. We are interested in testing whether the effect of academic probation changes monotonically with a student's high school grade percentile ranking ($X$).

Figure 3 provides a graphic summary of the data. The left graph shows close to but

Figure 3: Graphic Summary of Data




Note: Data are from Lindo et al. (2010). Local linear regressions are carried out using the triangular kernel and fixed bandwidth 0.6 as used in Lindo et al. (2010).

not perfect first-stage compliance. A closer look at the data finds that three out of a total of 44,362 students were put on academic probation even though their GPAs are above the cutoff for unknown reasons. Another 48 students with GPA below the cutoff were not put on academic probation. We can see from the data that 34 out of 48 students were suspended, likely for other reasons, but we can not find more details of the other 14 students. We drop those observations in our RD analysis. Including them has little impact on testing results reported below, but would make it more difficult to interpret the RD effects.

The middle and right graphs plot the conditional probability of a student leaving college after the first year (conditioning on a student's first-year GPA) for two separate subsamples determined by students' high school grades. For both subsamples, the estimated probability of leaving school drops at the probation cutoff, although data of the conditional probabilities are quite noisy. The size of the drop is larger for students whose high school grade is above the median.

Table 7: P-values of Proposed Monotonicity Tests

| | CATE Weakly Increases with X | | | | | | CATE Weakly Increases with X | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LFC | | | GMS | | | LFC | | | GMS | | |
| k | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 | 4.25 | 4.5 | 4.75 |
| Panel A: No Control Variable | | | | | | | | | | | | |
| $Q = 30$ | 0.34 | 0.24 | 0.22 | 0.34 | 0.24 | 0.22 | 0.21 | 0.17 | 0.14 | 0.21 | 0.16 | 0.13 |
| $Q = 40$ | 0.34 | 0.24 | 0.22 | 0.34 | 0.24 | 0.22 | 0.22 | 0.17 | 0.14 | 0.21 | 0.16 | 0.13 |
| Panel B: Control Variable ($S$) is a Dummy Variable for Gender | | | | | | | | | | | | |
| $Q = 30$ | 0.60 | 0.47 | 0.44 | 0.59 | 0.46 | 0.44 | 0.43 | 0.38 | 0.33 | 0.42 | 0.37 | 0.33 |
| $Q = 40$ | 0.61 | 0.51 | 0.47 | 0.61 | 0.50 | 0.46 | 0.47 | 0.39 | 0.34 | 0.45 | 0.37 | 0.33 |
| Panel C: Control Variable $S$ is a Dummy Variable for Native English Speaker | | | | | | | | | | | | |
| $Q = 30$ | 0.75 | 0.65 | 0.62 | 0.75 | 0.64 | 0.60 | 0.59 | 0.51 | 0.47 | 0.58 | 0.50 | 0.46 |
| $Q = 40$ | 0.83 | 0.73 | 0.77 | 0.83 | 0.72 | 0.69 | 0.68 | 0.59 | 0.55 | 0.66 | 0.59 | 0.55 |

Notes: Data are from Lindo et al. (2010). Nonparametric test statistics are calculated using the triangular kernel, the undersmoothed CCT bandwidth defined in Section 3.4, and the cubes defined in equation (3.6). All simulated critical values are calculated with 1,000 bootstrap repetitions. The interaction term method is defined in Section 5.

Table 7 reports the p-values of the proposed monotonicity tests. Again, the test is carried out using the triangular kernel, undersmoothed CCT bandwidth, and moment transformation detailed in the implementation procedure in Section 3.4. $Q = 20$ is set to 20 and 30 in this application following suggestions in Andrews and Shi (2013, 2014) as is discussed in Section 3.4. When $Q = 30$, the smallest effective sample size in kernel regressions used in our proposed testing procedure averages around 60 in panel A when the test is carried out without using any additional control. The number is around 20 in panels B and C, when the test is carried out using an additional control ($S$) for gender (male vs. female) or native language (native English speaker vs. nonnative English speaker), respectively.

Lindo et al. (2010) compare students with above-median high school grades to those with below-median high school grades and find that the discouragement effect of academic probation "is greater for students that performed relatively better in high school (above the median of students entering the university)", as they find that the discouragement effect is not statistically significant for students with below-median high school grades but is statistically significant at the 10% level for students with above-median high school

grades. The subsampling approach in Lindo et al. (2010) uses the high school grade variable after dichotomizing it by its median and also does not adjust for multiple testing. Our nonparametric test, in contrast, uses full infomation of any continuous random variable and is asymptotically size correct. In this empirical example, our proposed test does not find evidence supporting the conclusion that the discouragement effect of academic probation changes monotonically with high school grades in any of the three panels in Table 7. The testing result is robust across all reported choices of bandwidths, $Q$, and the method (LFC or GMS) of obtaining simulated critical values.

# A Appendix A: Proofs of Lemmas in Sections 3.1 and 4.1

This section proves the Lemmas in Section 3 that illustrate the equivalence between the original null hypothesis in (3.1) and the transformed null hypothesis in (3.7).

**Proof of Lemma 3.1:** First, we prove that (i) implies (ii). For any $x_1 > x_2$ and any $q = 2, 3, \ldots$, such that $q \cdot x_1, q \cdot x_2 \in \{0, 1, 2, \cdots, q\}$, we have $x_1 \geq x_2 + 1/q$. By (i), this implies that $\lambda(x) \geq \lambda(x)$ for all $x \in [x_1, x_1 + 1/q]$ and $x' \in [x_2, x_2 + 1/q]$. Therefore, the weighted average of $\lambda(x)$ over $x \in [x_1, x_1 + 1/q]$ has to be greater than or equal to that of $\lambda(x)$ over $x' \in [x_2, x_2 + 1/q]$. Equivalently,

$$\frac{\int_{x_1}^{x_1+q^{-1}} \lambda(x) \cdot h(x)dx}{\int_{x_1}^{x_1+q^{-1}} h(x)dx} \geq \frac{\int_{x_2}^{x_2+q^{-1}} \lambda(x) \cdot h(x)dx}{\int_{x_2}^{x_2+q^{-1}} h(x)dx}.$$

We prove the inequality in (ii).

Second, we prove that (ii) implies (i) by contradiction. Suppose that $\lambda(x) < \lambda(x')$ for some $x > x'$. By continuity of $\lambda(x)$, there exist $x_u > x_l > x'_u > x'_l$ such that $\lambda(x) < \lambda(x')$ for all $x \in [x_l, x_u]$, $x' \in [x'_l, x'_u]$. In turn, we can find a $q$ large enough such that for some $x_1$ and $x_2$, $q \cdot x_1, q \cdot x_2 \in \{0, 1, 2, \cdots, q-1\}$, such that $[x_1, x_1 + 1/q] \subseteq [x_l, x_u]$ and $[x_2, x_2 + 1/q] \subseteq [x'_l, x'_u]$. Then the weighted average of $\lambda(x)$ over $x \in [x_1, x_1 + 1/q]$ has to be strictly less than that over $x' \in [x_2, x_2 + 1/x]$. That is,

$$\frac{\int_{x_1}^{x_1+q^{-1}} \lambda(x) \cdot h(x)dx}{\int_{x_1}^{x_1+q^{-1}} h(x)dx} < \frac{\int_{x_2}^{x_2+q^{-1}} \lambda(x) \cdot h(x)dx}{\int_{x_2}^{x_2+q^{-1}} h(x)dx}.$$

This completes our proof. □

**Proof of Lemma 3.2:** Since $CLATE(x)$ is continuous on $x$, we can set $\lambda(x)$ to $CLATE(x)$ and apply the results of Lemma 3.1. Let $\lambda(x) = CLATE(x) = E_P[(Y(1) - Y(0))(T(1) - T(0))|X = x, Z = c]/E_P[T(1) - T(0)|X = x, Z = c]$ and $h(x) = E_P[T(1) -$

$T(0)|X = x, Z = c] \cdot f_{X|Z=c}(x|c)$. Then,

$$
\begin{aligned}
&\int_{x_1}^{x_1+1/q} \lambda(x) \cdot h(x)dx \\
=&\int g_{x_1,q}(x)E_P[(Y(1) - Y(0))(T(1) - T(0))|X = x, Z = c] \cdot f_{X|Z=c}(x|c)dx \\
=&\int g_{x_1,q}(x)E_P[Y(1)T(1) + Y(0)(1 - T(1))|X = x, Z = c] \cdot f_{X|Z=c}(x|c)dx \\
&- \int g_{x_1,q}(x)E_P[Y(1)T(0) + Y(0)(1 - T(0))|X = x, Z = c] \cdot f_{X|Z=c}(x|c)dx \\
\overset{(a)}{=}&\lim_{z \searrow c} \int g_{x_1,q}(x)E_P[Y(1)T(1) + Y(0)(1 - T(1))|X = x, Z = z] \cdot f_{X|Z=z}(x|z)dx \\
&- \lim_{z \nearrow c} \int g_{x_1,q}(x)E_P[Y(1)T(0) + Y(0)(1 - T(0))|X = x, Z = z] \cdot f_{X|Z=z}(x|z)dx \\
=&\lim_{z \searrow c} \int g_{x_1,q}(x)E_P[Y(1)T + Y(0)(1 - T)|X, Z = z] \cdot f_{X|Z=z}(x|z)dx \\
&- \lim_{z \nearrow c} \int g_{x_1,q}(x)E_P[Y(1)T + Y(0)(1 - T)|X, Z = z] \cdot f_{X|Z=z}(x|z)dx \\
=&\lim_{z \searrow c} E_P[g_{x_1,q}(X)E_P[Y|X, Z = z]|Z = z] - \lim_{z \nearrow c} E_P[g_{x_1,q}(X)E_P[Y|X, Z = z]|Z = z] \\
=&\lim_{z \searrow c} E_P[g_{x_1,q}(X)Y|Z = z] - \lim_{z \nearrow c} E_P[g_{x_1,q}(X)Y|Z = z], \\
\equiv&\rho_P^{(1)}(\ell).
\end{aligned}
$$

Note that equality (a) holds from Assumptions 2.1 and 3.1. Specifically, the assumptions imply uniform continuity of $E_P[Y(1)T(1) + Y(0)(1 - T(1))|X = x, Z = z]f_{X|Z}(x|z)$ and $E_P[Y(1)T(0) + Y(0)(1 - T(0))|X = x, Z = z]f_{X|Z}(x|z)$ in $x$ and $z$ on $\mathcal{X} \times N_{\delta,z}(c)$, which further implies equality (a) holds, as uniform continuity implies, for example, that for all $\epsilon > 0$ there exists $\delta' > 0$ such that $\sup_{x \in \mathcal{X}} |E_P[Y(1)T(1) + Y(0)(1 - T(1))|X = x, Z = z]f_{X|Z}(x|z) - E_P[Y(1)T(1) + Y(0)(1 - T(1))|X = x, Z = c]f_{X|Z}(x|c)| \leq \epsilon$ for all $|z - c| < \delta'$. Other equalities above follow from definitions of potential outcome and potential treatment variables.

Similarly, we can show that

$$\int_{x_1}^{x_1+1/q} h(x)dx = \lim_{z \searrow c} E_P[g_{x_1,q}(X)T|Z = z] - \lim_{z \nearrow c} E_P[g_{x_1,q}(X)T|Z = z] \equiv \varrho_P^{(1)}(\ell),$$

and that the same results apply to integrals concerning $x_2$. Then applying the results in Lemma 3.1, we know that testing the null hypothesis $H_{0,FRD}$ is equivalent to testing

$$\rho_P^{(2)}(\ell)\varrho_P^{(1)}(\ell) - \rho_P^{(1)}(\ell)\varrho_P^{(2)}(\ell) \leq 0$$

for all $q = 2, 3, \cdots$, and $x_1 > x_2$ such that $q \cdot (x_1, x_2) \in \{0, 1, 2, \cdots, q-1\}^2$. $\square$

**Proof of Lemma 4.1:** To prove the lemma, we first state and prove the following equivalence result.

**Lemma A.1** *Let $\lambda(x, s)$ be a continuous function in $(x, s)$ on $\mathcal{X} \times \mathcal{S}$, and $0 < h(x, s) \leq M < \infty$ be a weight function. The following two statements are equivalent:*

*(i) $\lambda(x_1, s) \geq \lambda(x_2, s)$ whenever $x_1 \geq x_2$ for any $x_1, x_2 \in \mathcal{X}$ and $s \in \mathcal{S}$;*

*(ii) for any $q \in 2, 3, \cdots$, and $x_1 \geq x_2$ such that $x_1, x_2 \in \{0, 1/q, 2/q, \cdots, (q-1)/q\}^{d_x}$ and $s \in \{0, 1/(q-1), 2/(q-1), \cdots, (q-2)/(q-1)\}^{d_s}$,*

$$\frac{\int_{C_{x_1,q} \times C_{s,q}} \lambda(x, s) \cdot h(x, s) dx ds}{\int_{C_{x_1,q} \times C_{s,q}} h(x, s) dx ds} \geq \frac{\int_{C_{x_2,q} \times C_{s,q}} \lambda(x, s) \cdot h(x, s) dx ds}{\int_{C_{x_2,q} \times C_{s,q}} h(x, s) dx ds}.$$

**Proof:** For notational simplicity and without loss of generality, we prove the lemma for the case with $d_w = d_s = 1$. Proof for the higher dimension case follows the same idea but requires more complicated notation.

First, we prove that (i) implies (ii). For any $x_1 = x_2$, the inequality in (ii) holds trivially with equality. For any $x_1 > x_2$ and any $q = 1, 2, 3, \ldots$, such that $(q+1) \cdot x_1, (q+1) \cdot x_2 \in \{0, 1, 2, \cdots, q\}$, we have $x_1 \geq x_2 + 1/(q+1)$. By (i), this implies that $\lambda(x, s) \geq \lambda(x', s)$ for all $x \in [x_1, x_1 + 1/(q+1)]$ and $w' \in [x_2, x_2 + 1/(q+1)]$ and $s \in \mathcal{S}$. Therefore, the weighted average of $\lambda(x, s)$ over $x \in [x_1, x_1 + /(q+1)]$ and $s \in C_{s,q}$ has to be greater or equal to that of $\lambda(x', s)$ over $x' \in [x_2, x_2 + 1/(q+1)]$ and $s \in C_{s,q}$. This is the inequality in (ii).

Second, we prove that (ii) implies (i) by contradiction. Suppose that $\lambda(x, s') < \lambda(x', s')$ for some $s'$ and $x \geq x'$. By continuity of $\lambda(x, s)$, there exist $x_u > x_l > x'_u > x'_l$ and $s_u > s_l$ such that $\lambda(x, s') < \lambda(x', s')$ for all $x \in [x_l, x_u]$, $x' \in [x'_l, x'_u]$ and $s' \in [s_l, s_u]$. In turn, we can find a $q$ large enough such that for some $x_1$, $x_2$, and $s$, $(q+1) \cdot x_1, (q+1) \cdot x_2 \in \{0, 1, 2, \cdots, q\}$, and $q \cdot s \in \{0, 1, 2, \cdots, q-1\}$ so that $[x_1, x_1 + 1/(q+1)] \subseteq [x_l, x_u]$, $[x_2, x_2 + 1/(q+1)] \subseteq [x'_l, x'_u]$ and $[s, s + 1/q] \subseteq [s_l, s_u]$. Then the weighted average of $\lambda(x, s)$ over $x \in [x_1, x_1 + 1/(q+1)]$ and $s \in [s, s + 1/q]$ has to be strictly less than that of $\lambda(x', s)$ over $x' \in [x_2, x_2 + 1/(q+1)]$ and $s \in [s, s + 1/q]$. This contradicts the inequality in (ii). The lemma is hence proven. $\square$

Set $\lambda(x,s)$ to $CLATE(x,s)$. Then the statement (i) in the above lemma A.1 reduces to the null hypothesis $H'_{0,FRD}$. Let the weight function in (ii) be $h(x,s) = E_P[T(1) - T(0)|X = x, S = s, Z = c] \cdot f_{X,S|Z}(x,s|c)$ with $f_{X,S|Z}(x,s|z)$ denoting the probability density function of $(X,S)|Z = z$, the inequality in (ii) would reduce to the inequality described in the next Lemma. A formal proof would be similar to the proof of Lemma 3.2, given the above-described results of Lemma A.1. We omit the details.

# B Appendix B: Asymptotic Properties of the Benchmark Test

## B.1 Regularity Conditions

In this section, we formalize the asymptotic properties of the benchmark test discussed in Section 3.3. Let $f_z(z)$ and $f_{xz}(x,z)$ denote the marginal density function of $Z$, and the joint density of $X$ and $Z$, respectively. Let $\zeta_{P,+}(x,z) = E_P[Y|X = x, Z = z]$, $\sigma^2_{P,+}(x,z) = V_P(Y|X = x, Z = z)$ and $\varsigma_{P,+}(x,z) = E_P[T|X = x, Z = z]$ for $z \geq c$ and $\zeta_{P,-}(x,z) = E_P[Y|X = x, Z = z]$, $\sigma^2_{P,-}(x,z) = V_P(Y|X = x, Z = z)$ and $\varsigma_{P,-}(x,z) = E_P[T|X = x, Z = z]$ for $z < c$. Let $\mathcal{N}^+_{\delta,z}(c) = \{z : \ 0 \leq z - c \leq \delta\}$ be a neighborhood of $Z$ from the cut-off value $c$ to the right and $\mathcal{N}^-_{\delta,z}(c) = \{z : \ 0 < c - z \leq \delta\}$ be a neighborhood from $c$ to the left. Let $P_z$ denote the distribution of $Z$ under $P$, and let $\mathcal{P}$ denote the collection of distributions $P$. We make the following assumptions.

**Assumption B.1** *Assume that for some $\delta > 0$ and all $P \in \mathcal{P}$, the following conditions are satisfied.*

*(i) The random variable $Z$ has the same distribution across all $P \in \mathcal{P}$, and its density $f_z(z)$ is uniformly bounded away from zero and twice continuously differentiable in $z$ on $\mathcal{N}_{\delta,z}(c)$.*

*(ii) $f_{xz}(x,z)$ is twice continuously differentiable in $z$ on $\mathcal{N}_{\delta,z}(c)$ for all $x \in \mathcal{X}_c$, and $\partial^2 f_{xz}(x,z)/\partial x \partial z$ is uniformly bounded on $\in \mathcal{X}_c \times \mathcal{N}_{\delta,z}(c)$.*

*(iii) for all $x \in \mathcal{X}_c$, $\zeta_{P,+}(x,z)$, $\varsigma_{P,+}(x,z)$, $\zeta_{P,-}(x,z)$, and $\varsigma_{P,-}(x,z)$ are all twice continuously differentiable in $z$ on $\mathcal{N}^+_{\delta,z}(c)$.*

*(iv)* $\partial\zeta_{P,+}(x,z)/\partial z$, $\partial^2\zeta_{P,+}(x,z)/\partial x\partial z$, $\partial\varsigma_{P,+}(x,z)/\partial z$, $\partial^2\varsigma_{P,+}(x,z)/\partial x\partial z$, $\partial z\zeta_{P,-}(x,z)/\partial z$, $\partial^2\zeta_{P,-}(x,z)/\partial x\partial z$, $\partial\varsigma_{P,-}(x,z)/\partial z$, *and* $\partial^2\varsigma_{P,-}(x,z)/\partial x\partial z$ *are all uniformly bounded on* $\mathcal{X}_c \times \mathcal{N}^+_{\delta,z}(c)$.

*(v)* *Both* $\sigma^2_{P,+}(x,z)$ *and* $\sigma^2_{P,-}(x,z)$ *are uniformly bounded and uniformly bounded away from zero on* $\mathcal{X}_c \times \mathcal{N}^+_{\delta,z}(c)$.

*(vi)* $E_P[Y^4|Z=z]$ *is uniformly bounded for all* $z \in \mathcal{N}_{\delta,z}(c)$.

*(vii)* $E_P[T(1)-T(0)|Z=c]$ *is uniformly bounded away from zero.*

Assumptions B.1($i$)-($iii$) are standard in nonparametric estimation. Assumptions B.1($iv$) is needed to show that the bias terms of the $\hat{\nu}(\ell)$ are asymptotically negligible uniformly over $\ell \in \mathcal{L}$ and $P \in \mathcal{P}$. Assumption B.1($v$) and ($vi$) are required for the covariance estimator of the limiting process to be uniformly consistent, which is in turn used for showing the validity of the multiplier bootstrap. Similar conditions are also assumed in Andrews and Shi (2014), Hsu (2016) and Hsu and Shen (2019). Assumptions B.1($vii$) is assumed such that the group of compliers which is the subpopulation of interest under fuzzy design is not of mass zero. Assumption B.1($v$) and ($vii$) imply that the asymptotic limit of $\hat{\sigma}^2_{\mu,\epsilon}(\ell)$ defined in Section 3.2 is bounded away from zero for all $\ell \in \mathcal{L}$.

**Assumption B.2** *Assume that*

*(i)* $K(\cdot)$ *is a non-negative symmetric bounded kernel with a compact support in* $R$, *and* $\int K(u)du = 1$;

*(ii)* $h \to 0$, $nh \to \infty$ *and* $nh^5 \to 0$ *as* $n \to \infty$.

Assumption B.2(i) is a standard kernel undersmoothing bandwidth condition for local linear estimation. Undersmoothing helps to eliminate the nuisance bias term and obtain centered asymptotic distributions of the local linear estimators.

**Assumption B.3** $\{U_i : 1 \leq i \leq n\}$ *is a sequence of i.i.d. random variables that is independent of the sample path of* $\{(Y_i, X_i, Z_i, T_i) : 1 \leq i \leq n\}$ *such that* $E[U_i] = 0$, $E[U_i^2] = 1$, *and* $E[|U_i|^4] < M$ *for some* $M > 0$.

**Assumption B.4** *Assume that*

*(i)* $a_n$ *is a sequence of non-negative numbers satisfying the conditions that* $\lim_{n\to\infty} a_n = \infty$ *and* $\lim_{n\to\infty} a_n/\sqrt{nh} = 0$*;*

*(ii)* $B_n$ *is a sequence of non-negative numbers satisfying the conditions that* $B_n$ *is non-decreasing,* $\lim_{n\to\infty} B_n = \infty$*, and* $\lim_{n\to\infty} B_n/a_n = 0$*.*

Assumptions B.3 and B.4 are stated in the main text. We just list them here for completeness.

## B.2 Uniform Size Control and Test Consistency

Let $\mathcal{P}$ be a set of DGPs that satisfies regularity conditions defined in Assumption B.1. Let $h_{1,P}(.) = (-\varrho_P^{(2)}(.), \varrho_P^{(1)}(.), \rho_P^{(2)}(.), -\rho_P^{(1)}(.))$ be a $1 \times 4$ vector of functions. Let $\ddot{m}(.) = (m^{(1)}(Y,X,.), m^{(2)}(Y,X,.), q^{(1)}(T,X,.), q^{(2)}(T,X,.))'$ be a $4 \times 1$ random vector and $h_{2,P}^+(\ell_1,\ell_2) = \lim_{z\searrow c} Cov_P(\ddot{m}(\ell_1), \ddot{m}(\ell_2)|Z=z)$, $h_{2,P}^-(\ell_1,\ell_2) = \lim_{z\nearrow c} Cov_P(\ddot{m}(\ell_1), \ddot{m}(\ell_2)|Z=z)$ be the left and right limits of its conditional variance-covariance matrix at $Z=c$.

Let $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2^+ \times \mathcal{H}_2^-$, where $\mathcal{H}_1 = \{h_{1,P}(\cdot) : P \in \mathcal{P}\}$, $\mathcal{H}_2^+ = \{h_{2,P}^+(\cdot,\cdot) : P \in \mathcal{P}\}$, and $\mathcal{H}_2^- = \{h_{2,P}^-(\cdot,\cdot) : P \in \mathcal{P}\}$. For any two functions $h = (h_1, h_2^+, h_2^-)$ and $\tilde{h} = (\tilde{h}_1, \tilde{h}_2^+, \tilde{h}_2^-)$ in the space of $\mathcal{H}$, define metric $d$ as

$$d(h,\tilde{h}) = \max\left\{ d_1(h_1,\tilde{h}_1), d_2(h_2^+,\tilde{h}_2^+), d_2(h_2^-,\tilde{h}_2^-) \right\},$$

where $d_1(h_1,\tilde{h}_1) = \sup_{\ell\in\mathcal{L}} \left\| h_1(\ell) - \tilde{h}_1(\ell) \right\|$, $d_2(h_2,\tilde{h}_2) = \sup_{\ell_1,\ell_2\in\mathcal{L}} \left\| h_2(\ell_1,\ell_2) - \tilde{h}_2(\ell_1,\ell_2) \right\|$ and $\|\cdot\|$ is the Euclidean norm.

Let $\mathcal{P}_0$ be the subset of $\mathcal{P}$ such that the null hypothesis in (3.1) holds, and $\mathcal{L}^o(P) = \{\ell : \mu_P(\ell) = 0\}$ be the collection of $\ell$ indices satisfying the LFC of $H_{0,FRD}^T$ in (3.7) under $P$. Then we have the following results of the proposed monotonicity test.

**Theorem B.1** *Suppose that Assumptions 2.1, 3.1, and Assumptions B.1, B.2, B.3 and B.4 stated in Appendix B hold. Then, for every compact subset* $\mathcal{H}_{cpt}$ *of* $\mathcal{H}$*, we have*

*(a)* $\limsup_{n\to\infty} \sup_{\{P\in\mathcal{P}_0:\ h_P\in\mathcal{H}_{cpt}\}} P(\widehat{T}_{n,FRD} > \hat{c}_{n,FRD}^{\eta,LFC}(\alpha)) \le \alpha$*;*

*(b)* $\limsup_{n\to\infty} \sup_{\{P\in\mathcal{P}_0:\ h_P\in\mathcal{H}_{cpt}\}} P(\widehat{T}_{n,FRD} > \hat{c}^{\eta,GMS}_{n,FRD}(\alpha)) \leq \alpha$*;*

*(c) if there exists some* $P_c^{LFC} \in \mathcal{P}_0$ *such that* $\mathcal{L}^o(P_c^{LFC}) = \mathcal{L}$ *and the covariance kernel* $h_{2,\mu,P_c^{LFC}}(\ell_1,\ell_2) \equiv h_{1,P_c^{LFC}}(\ell_1)\left(h^+_{2,P_c^{LFC}}(\ell_1,\ell_2) + h^-_{2,P_c^{LFC}}(\ell_1,\ell_2)\right) h_{1,P_c^{LFC}}(\ell_2)'$ *under* $P_c^{LFC}$ *is not a zero function, then*

$$\lim_{\eta\to 0}\limsup_{n\to\infty} \sup_{\{P\in\mathcal{P}_0:\ h_P\in\mathcal{H}_{cpt}\}} P(\widehat{T}_{n,FRD} > \hat{c}^{\eta,LFC}_{n,FRD}(\alpha)) = \alpha;$$

*(d) if there exists some* $P_c \in \mathcal{P}_0$ *such that* $\mathcal{L}^o(P_c)$ *is not empty and the covariance kernel* $h_{2,\mu,P_c}(\ell_1,\ell_2) \equiv h_{1,P_c}(\ell_1)\left(h^+_{2,P_c}(\ell_1,\ell_2) + h^-_{2,P_c}(\ell_1,\ell_2)\right) h_{1,P_c}(\ell_2)'$ *under* $P_c$ *is not a zero function, then*

$$\lim_{\eta\to 0}\limsup_{n\to\infty} \sup_{\{P\in\mathcal{P}_0:\ h_P\in\mathcal{H}_{cpt}\}} P(\widehat{T}_{n,FRD} > \hat{c}^{\eta,GMS}_{n,FRD}(\alpha)) = \alpha.$$

Let $P^*$ be an element of $\mathcal{P}$ such that there exists some $x_1^*, x_2^* \in \mathcal{X}$ with $x_2^* > x_1^*$ and $CLATE(x_2^*) < CLATE(x_1^*)$. The next theorem shows the consistency property of the proposed benchmark test under the same set of assumptions required in Theorem B.1.

**Theorem B.2** *Suppose that Assumptions 2.1, 3.1, and Assumptions B.1, B.2, B.3 and B.4 stated in Appendix B hold, then*

*(a)* $\lim_{n\to\infty} P^*(\widehat{T}_{n,FRD} > \hat{c}^{\eta,LFC}_{n,FRD}(\alpha)) = 1.$

*(b)* $\lim_{n\to\infty} P^*(\widehat{T}_{n,FRD} > \hat{c}^{\eta,GMS}_{n,FRD}(\alpha)) = 1.$

## B.3 Proof of Thoerems B.1 and B.2

To prove the asymptotic properties of the proposed test described in Theorems B.1 and B.1, we first give out auxiliary lemma that will be used in the main proof of the theorems.

Let $E_Z$ denote the expectation conditional on sample path $\{Z_1, Z_2, \ldots\}$. Define $C_k = \frac{\int_0^\infty (\vartheta_2 - u\vartheta_1)^2 K^2(u)du}{(\vartheta_2\vartheta_0 - \vartheta_1^2)^2 \cdot f_z(c)}$. And let $\{P_n : n \geq 1\}$ be a sequence of distributions.

**Lemma B.1** *Suppose* $P_n \in \mathcal{P}$ *for all* $n$ *so each of the distributions satisfies Assumption B.1. Suppose Assumption B.2 also holds. Then for any subsequence* $\{P_{k_n}\}$ *of* $\{P_n\}$ *such*

*that* $\lim_{n\to\infty} d_2(h^+_{2,P_{k_n}}, h_2^{*+}) = 0$ *for some* $h_2^{*+} \in \mathcal{H}_2^+$ *and* $\lim_{n\to\infty} d_2(h^-_{2,P_{k_n}}, h_2^{*-}) = 0$ *for some* $h_2^{*-} \in \mathcal{H}_2^-$*, we have*

$$\sqrt{k_n h}\begin{pmatrix} \hat{m}^{(1)}_{k_n,+} - m^{(1)}_{P_{k_n},+} \\ \hat{m}^{(2)}_{k_n,+} - m^{(2)}_{P_{k_n},+} \\ \hat{q}^{(1)}_{k_n,+} - q^{(1)}_{P_{k_n},+} \\ \hat{q}^{(2)}_{k_n,+} - q^{(2)}_{P_{k_n},+} \end{pmatrix} \Rightarrow \Phi_{C_k h_2^{*+}} \ , \quad \sqrt{k_n h}\begin{pmatrix} \hat{m}^{(1)}_{k_n,-} - m^{(1)}_{P_{k_n},-} \\ \hat{m}^{(2)}_{k_n,-} - m^{(2)}_{P_{k_n},-} \\ \hat{q}^{(1)}_{k_n,-} - q^{(1)}_{P_{k_n},-} \\ \hat{q}^{(2)}_{k_n,-} - q^{(2)}_{P_{k_n},-} \end{pmatrix} \Rightarrow \Phi_{C_k h_2^{*-}} \ ,$$

*where* $\Phi_{C_k h_2^{*+}}$ *and* $\Phi_{C_k h_2^{*-}}$ *are independent mean zero Gaussian processes with covariance kernels* $C_k h_2^{*+}$ *and* $C_k h_2^{*-}$*.*

**Proof of Lemma B.1:** Here we prove the first weak convergence result. The second one follows by essentially the same proof but with subscripts + replaced by subscripts −. The resulting two limiting Gaussian processes are independent because the local linear estimators involved in the two convergence results use different subsamples and the data are independent.

By the Cramér-Wold Theorem, it is sufficient to show the $m^{(1)}_{P_{k_n},+}(\ell)$ case. Note that

$$\begin{aligned}
&\sqrt{k_n h}(\hat{m}^{(1)}_{k_n,+}(\ell) - m^{(1)}_{P_{k_n},+}(\ell)) \\
=&\sum_{i=1}^{k_n} \sqrt{k_n h}(\mathsf{w}^+_{k_n i}(m^{(1)}(Y_i, X_i, \ell) - m^{(1)}_{P_{k_n},+}(\ell))) \\
=&\sum_{i=1}^{k_n} \sqrt{k_n h}(\mathsf{w}^+_{k_n i}(m^{(1)}(Y_i, X_i, \ell) - E_Z[m^{(1)}(Y_i, X_i, \ell)])) \\
&+ \sum_{i=1}^{k_n} \sqrt{k_n h}(\mathsf{w}^+_{k_n i}(E_Z[m^{(1)}(Y_i, X_i, \ell)] - m^{(1)}_{P_{k_n},+}(\ell))).
\end{aligned}$$

We first consider the second term, which is the bias term. By Theorem 4 of Fan and Gijbels (1992), we know that

$$\sum_{i=1}^{k_n} \sqrt{k_n h}(\mathsf{w}^+_{k_n i}(E_Z[m^{(1)}(Y_i, X_i, \ell)] - m^{(1)}_{P_{k_n},+}(\ell))) = O_p(\sqrt{k_n h^5}) = o_p(1).$$

Note that the first equality holds because the magnitude of the bias is proportional to the second derivative of $m^{(1)}_{P_{k_n},+}(\ell)$ with respect to $z$. By Assumption B.1, we know that for all $P_{k_n}$, $\partial^2 \zeta_{P_{k_n},+}(x,z)/\partial z \partial z$ is uniformly bounded on $x \in \mathcal{X}_c$ and $z \in \mathcal{N}^+_{\delta,z}(c)$. Since $m^{(1)}_{P_{k_n},+}(\ell) = \lim_{z \searrow c} E_{P_{k_n}}[g_{x_1,q}(X)Y | Z = z]$, $m^{(1)}_{P_{k_n},+}(\ell)$ is uniformly bounded as

well. Then given the additional assumption that $k_n h^5 \to 0$, we know that the above $o_p(1)$ result holds uniformly over $\ell \in \mathcal{L}$.

Therefore, uniformly over $\ell \in \mathcal{L}$, we have

$$\begin{aligned}
&\sqrt{k_n h}(\hat{m}^{(1)}_{k_n,+}(\ell) - m^{(1)}_{P_{k_n},+}(\ell)) \\
=&\sum_{i=1}^{k_n} \sqrt{k_n h}(\mathsf{w}^+_{k_n i}(m^{(1)}(Y_i, X_i, \ell) - E_Z[m^{(1)}(Y_i, X_i, \ell)])) + o_p(1).
\end{aligned}$$

It is then easy to show that $\{(m^{(1)}(Y_i, X_i, \ell) : \ 1 \le i \le k_n, \ n \ge 1\}$ satisfies the manageability condition in the functional central limit theorem (FCLT), or Theorem 10.6 of Pollard (1990). The arguments are similar to those in the proof of Lemma 3.2 of Hsu and Shen (2019) and hold along the sequence $\{P_{k_n} : \ n \ge 1\}$. □

**Lemma B.2** *Suppose $P_n \in \mathcal{P}$ for all $n$ so each of the distribution satisfies Assumption B.1. Suppose B.2 also holds. Then for any subsequence $\{P_{k_n}\}$ of $\{P_n\}$ such that for $\lim_{n\to\infty} d(h_{P_{k_n}}, h^*) = 0$ for some $h^* \in \mathcal{H}$, we have*

$$\begin{aligned}
&\sup_{\ell\in\mathcal{L}} \left| \sqrt{k_n h}(\hat{\mu}_{k_n}(\ell) - \mu_{P_{k_n}}(\ell)) - \sum_{i=1}^{k_n} \phi_{\mu,k_n i}(\ell) \right| = o_p(1), \ \textit{and} \\
&\sqrt{k_n h}\left(\hat{\mu}_{k_n} - \mu_{P_{k_n}}\right) \Rightarrow \Phi_{C_k h^*_{2,\mu}}
\end{aligned}$$

*where $h^*_{2,\mu} = h^*_1(h^{*+}_2 + h^{*-}_2)h^{*\prime}_1$.*

**Proof of Lemma B.2:** First, note that Lemma B.1 implies that for $\kappa = 1, 2$,

$$\begin{aligned}
&\sup_{\ell\in\mathcal{L}} |\hat{\rho}^{(\kappa)}_{k_n}(\ell) - \rho^{(\kappa)}_{P_{k_n}}(\ell)| = O_p((k_n h)^{-1/2}), \\
&\sup_{\ell\in\mathcal{L}} |\hat{\varrho}^{(\kappa)}_{k_n}(\ell) - \varrho^{(\kappa)}_{P_{k_n}}(\ell)| = O_p((k_n h)^{-1/2}).
\end{aligned} \tag{B.1}$$

Next, note that

$$\begin{aligned}
&\sqrt{k_n h}(\hat{\rho}^{(2)}_{k_n}(\ell)\hat{\varrho}^{(1)}_{k_n}(\ell) - \rho^{(2)}_{P_{k_n}}(\ell)\varrho^{(1)}_{P_{k_n}}(\ell)) \\
=&\hat{\varrho}^{(1)}_{k_n}(\ell)\sqrt{k_n h}\big(\hat{\rho}^{(2)}_{k_n}(\ell) - \rho^{(2)}_{P_{k_n}}(\ell)\big) - \rho^{(2)}_{P_{k_n}}(\ell)\sqrt{k_n h}\big(\hat{\varrho}^{(1)}_{k_n}(\ell) - \varrho^{(1)}_{P_{k_n}}(\ell)\big) \\
=&\varrho^{(1)}_{P_{k_n}}(\ell)\sqrt{k_n h}\big(\hat{\rho}^{(2)}_{k_n}(\ell) - \rho^{(2)}_{P_{k_n}}(\ell)\big) - \rho^{(2)}_{P_{k_n}}(\ell)\sqrt{k_n h}\big(\hat{\varrho}^{(1)}_{k_n}(\ell) - \varrho^{(1)}_{P_{k_n}}(\ell))\big) \\
&+ \sqrt{k_n h}\big(\hat{\rho}^{(2)}_{k_n}(\ell) - \rho^{(2)}_{P_{k_n}}(\ell)\big)\big(\hat{\varrho}^{(1)}_{k_n}(\ell) - \varrho^{(1)}_{P_{k_n}}(\ell))\big) \\
=&\varrho^{(1)}_{P_{k_n}}(\ell)\sqrt{k_n h}\big(\hat{\rho}^{(2)}_{k_n}(\ell) - \rho^{(2)}_{P_{k_n}}(\ell)\big) - \rho^{(2)}_{P_{k_n}}(\ell)\sqrt{k_n h}\big(\hat{\varrho}^{(1)}_{k_n}(\ell) - \varrho^{(1)}_{P_{k_n}}(\ell))\big) + o_p(1)
\end{aligned}$$

where the $o_p(1)$ result holds uniformly over $\ell \in \mathcal{L}$ due to Equation (B.1). Further, since

$$\sqrt{k_n h}(\hat{\rho}^{(\kappa)}_{k_n}(\ell) - \rho^{(\kappa)}_{P_{k_n}}(\ell)) = \sum_{i=1}^{k_n} \phi^{(\kappa)}_{\rho,k_n i}(\ell), \qquad \sqrt{k_n h}(\hat{\varrho}^{(\kappa)}_{k_n}(\ell) - \varrho^{(\kappa)}_{P_{k_n}}(\ell)) = \sum_{i=1}^{k_n} \phi^{(\kappa)}_{\varrho,k_n i}(\ell),$$

for all $\ell \in \mathcal{L}$, we have that

$$\begin{aligned}
&\sqrt{k_n h}(\hat{\rho}^{(2)}_{k_n}(\ell)\hat{\varrho}^{(1)}_{k_n}(\ell) - \rho^{(2)}_{P_{k_n}}(\ell)\varrho^{(1)}_{P_{k_n}}(\ell)) \\
=&\sum_{i=1}^{k_n} \varrho^{(1)}_{P_{k_n}}(\ell)\phi^{(2)}_{\rho,k_n i}(\ell) - \sum_{i=1}^{k_n} \rho^{(2)}_{P_{k_n}}(\ell)\phi^{(1)}_{\varrho,k_n i}(\ell) + o_p(1).
\end{aligned}$$

and the $o_p(1)$ result holds uniformly over $\ell \in \mathcal{L}$.

Similarly, we can write

$$\sqrt{k_n h}(\hat{\rho}^{(1)}_{k_n}(\ell)\hat{\varrho}^{(2)}_{k_n}(\ell) - \rho^{(1)}_{P_{k_n}}(\ell)\varrho^{(2)}_{P_{k_n}}(\ell)) = \sum_{i=1}^{k_n} \varrho^{(2)}_{P_{k_n}}(\ell)\phi^{(1)}_{\rho,k_n i}(\ell) - \sum_{i=1}^{k_n} \rho^{(1)}_{P_{k_n}}(\ell)\phi^{(2)}_{\varrho,k_n i}(\ell) + o_p(1).$$

Finally, we have

$$\begin{aligned}
&\sqrt{k_n h}(\hat{\mu}_{k_n}(\ell) - \mu_{P_{k_n}}(\ell)) \\
=&\sqrt{k_n h}(\hat{\rho}^{(2)}_{k_n}(\ell)\hat{\varrho}^{(1)}_{k_n}(\ell) - \rho^{(2)}_{P_{k_n}}(\ell)\varrho^{(1)}_{P_{k_n}}(\ell) - \hat{\rho}^{(1)}_{k_n}(\ell)\hat{\varrho}^{(2)}_{k_n}(\ell) + \rho^{(1)}_{P_{k_n}}(\ell)\varrho^{(2)}_{P_{k_n}}(\ell)) \\
=&\sum_{i=1}^{k_n} \varrho^{(1)}_{P_{k_n}}(\ell)\phi^{(2)}_{\rho,k_n i}(\ell) - \rho^{(2)}_{P_{k_n}}(\ell)\phi^{(1)}_{\varrho,k_n i}(\ell) - \varrho^{(2)}_{P_{k_n}}(\ell)\phi^{(1)}_{\rho,k_n i}(\ell) + \rho^{(1)}_{P_{k_n}}(\ell)\phi^{(2)}_{\varrho,k_n i}(\ell) + o_p(1) \\
=&\sum_{i=1}^{k_n} \phi_{\mu,k_n i}(\ell) + o_p(1).
\end{aligned}$$

with the $o_p(1)$ result holding uniformly over $\ell \in \mathcal{L}$.

The above equation shows the first part of Lemma B.2. To show the second part, we will apply the Theorem 10.6 of Pollard (1990). Define our triangular array as $\{\phi_{\mu,k_n i}(\ell) : \ \ell \in \mathcal{L},\ i \leq k_n,\ 1 \leq n\}$. Note that by the same argument as in Lemma A1 of Hsu et al. (2019), we can show that the triangular is manageable so the part (i) of Theorem 10.2 of Pollard (1990) holds. We can apply similar arguments of Lemma 3.2 of Hsu and Shen (2019) to show that Parts (ii)-(v) hold too. These would complete our proof and we omit the details for brevity. □

**Lemma B.3** *Assume that Assumptions B.1, B.2, and B.3 hold. For any subsequence of $k_n$ of $n$ such that $\lim_{n\to\infty} d(h_{P_{k_n}}, h^*) = 0$ for some $h^* \in \mathcal{H}$, we have the simulated process $\widehat{\Phi}^u_{\mu,k_n}(\cdot) \Rightarrow \Phi_{C_k h^*_{2,\mu}}(\cdot)$ conditional on sample path with probability approaching 1.*

**Proof of Lemma B.3:** Recall that $\widehat{\Phi}^u_{\mu,k_n}(\cdot) = \sum_{i=1}^{k_n} U_i \cdot \hat{\phi}_{\mu,k_n i}(\cdot)$. It is straightforward to see that $\{U_i \cdot \hat{\phi}_{\mu,k_n i}(\ell) : \ \ell \in \mathcal{L},\ i \leq k_n,\ 1 \leq n\}$ is manageable. Define $\ddot{h}_{2,k_n,\mu}(\ell_1,\ell_2) = \sum_{i=1}^{k_n} \hat{\phi}_{\mu,k_n i}(\ell_1)\hat{\phi}_{\mu,k_n i}(\ell_2)$. We know that $\sup_{\ell_1,\ell_2\in\mathcal{L}} |\ddot{h}_{2,k_n,\mu}(\ell_1,\ell_2) - C_k h^*_{2,\mu}(\ell_1,\ell_2)| \xrightarrow{p} 0$. The rest of the proof is similar to that for Lemma 3.3 of Hsu and Shen (2019) and we omit the details. □

**Lemma B.4** *Let* $\hat{\sigma}^2_{\mu,k_n,\epsilon}(\ell) = \max\{\hat{\sigma}^2_{\mu,k_n}(\ell), \epsilon \cdot \hat{\sigma}^2_{\mu,k_n}(\ell_0)\}$ . *Assume that Assumptions B.1 and B.2 hold. For any subsequence of* $k_n$ *of* $n$ *such that* $\lim_{n\to\infty} d(h_{P_{k_n}}, h^*) = 0$ *for some* $h^* \in \mathcal{H}$*, we know that* $\sup_{\ell\in\mathcal{L}} |\hat{\sigma}_{\mu,k_n,\epsilon}(\ell)^{-1} - \sigma^*_{\mu,\epsilon}(\ell)^{-1}| \xrightarrow{p} 0$*, where* $\sigma^*_{\mu,\epsilon}(\ell) = \max\{\big(C_k h^*_{2,\mu}(\ell,\ell)\big)^{1/2}, \big(\epsilon \cdot C_k h^*_{2,\mu}(\ell_0,\ell_0)\big)^{1/2}\}$.

**Proof of Lemma B.4:** Since $\hat{\sigma}_{\mu,k_n}(\ell) = \ddot{h}_{2,k_n,\mu}(\ell,\ell)$, where $\ddot{h}_{2,k_n,\mu}(\ell,\ell)$ is defined in the proof of Lemma B.3, we know that $\sup_{\ell\in\mathcal{L}} |\hat{\sigma}_{\mu,k_n}(\ell) - \big(C_k h^*_{2,\mu}(\ell,\ell)\big)^{1/2}| \xrightarrow{p} 0$. Next, by the fact that the maximum operator is a continuous functional, we have $\sup_{\ell\in\mathcal{L}} |\hat{\sigma}_{\mu,k_n,\epsilon}(\ell) - \sigma^*_{\mu,\epsilon}(\ell)| \xrightarrow{p} 0$. Further, since $\sigma^*_\mu(\ell_0)$ is bounded away from zero under Assumption B.1, it follows that $\sup_{\ell\in\mathcal{L}} |\hat{\sigma}_{\mu,\epsilon}(\ell)^{-1} - \sigma^*_{\mu,\epsilon}(\ell)^{-1}| \xrightarrow{p} 0$. This completes the proof of Lemma B.4. □

**Proof of Theorem B.1:** Note that by construction, $\hat{c}^{\eta,LFC}_{n,FRD}(\alpha) \geq \hat{c}^{\eta,GMS}_{n,FRD}(\alpha)$, so the size of the test based on the LFC critical value is always smaller than that based on $\hat{c}^{\eta,GMS}_{n,FRD}(\alpha)$. Therefore, it is sufficient to show that the test based on $\hat{c}^{\eta,GMS}_{n,FRD}(\alpha)$ has uniform size control. To show this, we can apply the same arguments as in the proof of Theorem 4.1 of Hsu et al. (2019) given Lemmas B.2, B.3 and B.4, and we omit the details.

To show part (d) of the theorem, note that if there exists $P_c \in \mathcal{P}_0$ such that $\mathcal{L}^o(P_c)$ is not empty and $h_{2,\mu,P_c}$ restricted to $\mathcal{L}^o(P_c) \times \mathcal{L}^o(P_c)$ is not a zero function, then by the same proof based on the pointwise asymptotics as in Lemma 1 of Donald and Hsu (2016) and by Tsirel'Son (1976), we have that under $P_c$, the CDF function, $G(.)$, of the limiting null distribution of $\widehat{T}_{n,FRD}$ is continuous and is strictly increasing on $(0,\infty)$, and $G(0) > 1/2$. Then, by the same proof for Theorem 2(b) of Andrews and Shi (2013), it is true that under $P_c$, $\lim_{\eta\to 0} P(\widehat{T}_{n,FRD} > \hat{c}^{\eta,GMS}_{n,FRD}(\alpha)) = \alpha$, which implies

that $\lim_{\eta\to 0}\limsup_{n\to\infty}\sup_{\{P\in\mathcal{P}_0:\ h_P\in\mathcal{H}_{cpt}\}} P(\widehat{T}_{n,FRD} > \hat{c}^{\eta,GMS}_{n,FRD}(\alpha)) \geq \alpha$. Then by combining the result in part (b) of the Theorem, we obtain the uniform size control result in part (d).

To show part (c) of the theorem, the asymptotic results in Lemmas B.2, B.3 and B.4 directly imply that $\lim_{\eta\to 0} P_c^{LFC}(\widehat{T}_{n,FRD} > \hat{c}^{\eta,LFC}_{n,FRD}(\alpha)) = \alpha$ under $P_c^{LFC}$. Then we know $\lim_{\eta\to 0}\limsup_{n\to\infty}\sup_{\{P\in\mathcal{P}_0:\ h_P\in\mathcal{H}_{cpt}\}} P(\widehat{T}_{n,FRD} > \hat{c}^{\eta,LFC}_{n,FRD}(\alpha)) \geq \alpha$. Combining the result in part (a) of the theorem, we obtain the result in part (c). This completes our proof. □

**Proof of Theorem B.2:** Suppose $CLATE(x_2^*) > CLATE(x_1^*)$ for some $x_2^* < x_1^*$ and some $s$. Then by continuity of $CLATE(x)$, we can find $x_2' \ll x_1'$ such that $CLATE(x_2') > CLATE(x_1')$. Again, by continuity of $CLATE(x)$, we can find a small $\delta$ such that for all $x_2'' \in \mathcal{N}_{\delta,x}(x_2')$ and $x_1'' \in \mathcal{N}_{\delta,x}(x_1')$, we have $x_2'' \ll x_1''$ and $CLATE(x_2'') > CLATE(x_1'')$. Then we can find a $q$ large enough and $\ddot{\ell} = (\ddot{x}_1, \ddot{x}_2, q) \in \mathcal{L}$ such that $\Pi_{j=1}^{d_x}[\ddot{x}_{j1}, \ddot{x}_{j1}+1/(q+1)] \subseteq \mathcal{N}_{\delta,x}(x_1')$, and $\Pi_{j=1}^{d_x}[\ddot{x}_{j2}, \ddot{x}_{j2}+1/(q+1)] \subseteq \mathcal{N}_{\delta,x}(x_2')$. It is then straightforward to see that $\mu_{P^*}(\ddot{\ell}) > 0$.

By the definition of $\widehat{T}_{n,FRD}$, we know that $\widehat{T}_{n,FRD} \geq \sqrt{nh}\hat{\mu}_n(\ddot{\ell})/\hat{\sigma}_{\mu,\epsilon}(\ddot{\ell})$, and $\widehat{T}_{n,FRD}$ will diverge to positive infinity when $n \to \infty$, because $\sqrt{nh}\hat{\mu}_n(\ddot{\ell})$ will diverge to positive infinity and $\hat{\sigma}_{\mu,\epsilon}(\ddot{\ell})$ is bounded in probability. Also, both simulated critical values $\hat{c}^{\eta,LFC}_{n,FRD}(\alpha)$ and $\hat{c}^{\eta,GMS}_{n,FRD}(\alpha)$ are bounded in probability. The consistency result of the proposed monotonicity tests then follows. □

## References


Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97(457), 284–292.

Andrews, D. and G. Soares (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78, 119–157.

Andrews, D. W. and P. Guggenberger (2009): "Validity of subsampling and 'plug-in asymptotic' inference for parameters defined by moment inequalities," *Econometric Theory*, 25, 669–709.

Andrews, D. W. and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666.

——— (2014): "Nonparametric Inference Based on Conditional Moment Inequalities," *Journal of Econometrics*, 179, 31–45.

——— (2017): "Inference Based on Many Conditional Moment Inequalities," *Journal of Econometrics*, 196(2), 275–287.

Angrist, J. D. and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *The Quarterly Journal of Economics*, 114, 533–575.

Angrist, J. D. and M. Rokkanen (2015): "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff," *Journal of the American Statistical Association*, 110(512), 1331–1344.

Barone, G., F. DAcunto, and G. Narciso (2015): "Telecracy: Testing for Channels of Persuasion," *American Economic Journal: Economic Policy*, 7(2), 3060.

Bertanha, M. (2016): "Regression Discontinuity Design with Many Thresholds," *working paper*.

Bertanha, M. and G. W. Imbens (2014): "External validity in fuzzy regression discontinuity designs," *working paper, National Bureau of Economic Research*, No. w20773.

Black, S. (1999): "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, 114(2), 577–599.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2019): "Regression discontinuity designs using covariates," *Review of Economics and Statistics*, 101(3), 442–451.

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

Carneiro, P., K. V. Loken, and K. G. Salvanes (2015): "A Flying Start? Maternity Leave Benefits and Long-Run Outcomes of Children," *Journal of Political Economy*, 123(2), 365–412.

Cattaneo, M. D., R. Titiunik, G. Vazquez-Bare, and L. Keele (2016): "Interpreting Regression Discontinuity Designs with Multiple Cutoffs," *The Journal of Politics*, 78(4), 1229–1248.

Chetverikov, D. (2019): "Testing Regression Monotonicity in Econometric Models," *Econometric Theory*, 35(4), 729–776.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2008): "Nonparametric Tests for Treatment Effect Heterogeneity," *The Review of Economics and Statistics*, 90(3), 389405.

Donald, S. G. and Y.-C. Hsu (2016): "Improving the Power of Tests of Stochastic Dominance," *Econometric Review*, 35, 553–585.

Dong, Y. (2019): "Regression discontinuity designs with sample selection," *Journal of Business & Economic Statistics*, 37(1), 171–186.

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.

Fan, J. and I. Gijbels (1992): "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 20(4), 2008–2036.

Feir, D., T. Lemieux, and V. Marmer (2016): "Weak Identification in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 34(2), 185–196.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75, 259–276.

Frandsen, B. R., M. Frölich, and B. Melly (2012): "Quantile Treatment Effects in the Regression Discontinuity Design," *Journal of Econometrics*, 168, 382–395.

Frölich, M. and M. Huber (2019): "Including covariates in the regression discontinuity design," *Journal of Business & Economic Statistics*, 37(4), 736–748.

Ghosal, S., A. Sen, and A. W. van der Vaart (2000): "Testing Monotonicity of Regression," *Annals of Statistics*, 28(4), 1054–1082.

Hall, P. and N. E. Heckman (2000): "Testing for Monotonicity of a Regression Mean by Calibrating for Linear Functions," *Annals of Statistics*, 28(1), 20–39.

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

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *The Review of Economic Studies*, 65(2), 261–294.

Hotz, V., G. W.Imben, and J. H.Mortimerc (2005): "Predicting the efficacy of future training programs using past experiences at other locations," *Journal of Econometrics*, 125, 241–270.

Hsu, Y.-C. (2016): "Multiplier Bootstrap," *working paper*.

Hsu, Y.-C., C.-A. Liu, and X. Shi (2019): "Testing Generalized Regression Monotonicity," *Econometric Theory, forthcoming*.

Hsu, Y.-C. and S. Shen (2019): "Testing Treatment Effect Heterogeneity in Regression Discontinuity Designs," *Journal of Econometrics*, 208, 468–486.

Imbens, G. and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *The Review of Economic Studies*, 79, 933–959.

Ito, K. (2015): "Asymmetric Incentives in Subsidies: Evidence from a Large-Scale Electricity Rebate Program," *American Economic Journal: Economic Policy*, 7(3), 209–237.

Lee, D. S. and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 281–355.

Lindo, J. M., N. J. Sanders, and P. Oreopoulos (2010): "Ability, gender, and performance standards: Evidence from academic probation," *American Economic Journal: Applied Economicss*, 2, 95–117.

Linton, O., K. Song, and Y.-J. Whang (2010): "An improved bootstrap test of stochastic dominance," *Journal of Econometrics*, 154, 186–202.

Pollard, D. (1990): "Empirical Processes: Theory and Applications," in *NSF-CBMS Regional Conference Series in Probability and Statistics*.

Pop-Eleches, C. and M. Urquiola (2013): "Going to a Better School: Effects and Behavioral Responses," *American Economic Review*, 103, 1289–1324.

Shen, S. and X. Zhang (2016): "Distributional Test for Regression Discontinuity: Theory and Applications," *Review of Economics and Statistics, forthcoming*, 98, 685–700.

Tsirel'Son, V. S. (1976): "The density of the distribution of the maximum of a Gaussian process," *Theory of Probability & Its Applications*, 20(4), 847–856.

van der Klaauw, W. (2002): "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression Discontinuity Approach," *International Economic Review*, 43(4), 1249–1287.

Wager, S. and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113(523), 1228–1242.