Testing Treatment Effect Heterogeneity in Regression Discontinuity Designs

Yu-Chin Hsu[†]

Institute of Economics, Academia Sinica

Department of Finance, National Central University

Department of Economics, National Chengchi University

Shu Shen[‡]

Department of Economics, University of California, Davis

[†] E-mail: ychsu@econ.sinica.edu.tw. 128 Academia Road, Section 2, Nankang, Taipei, 115 Taiwan.

[‡] E-mail: shushen@ucdavis.edu. One Shields Avenue, Davis, CA 95616.

Acknowledgements: The authors would like to thank Arthur Lewbel, Vadim Marmer, Yuya Sasaki, Kevin Song, Liangjun Su, Ping Yu, and seminar participants at the department of economics in Havard/MIT, Singapore Management University, UC Irvine, University of British Columbia, University of Hong Kong, the department of statistics in UC Davis, the 2016 Australasia Meeting of the Econometric Society, the 2016 All California Econometrics Conference, the 2016 Annual Meeting of Taiwan Econometric Society, the 2016 Cross-Strait Dialogue IV Workshop and the 2016 Midwest Econometrics Group for helpful comments. Yu-Chin Hsu gratefully acknowledges research support from the Ministry of Science and Technology of Taiwan (MOST103-2628-H-001-001-MY4) and the Career Development Award of Academia Sinica, Taiwan. Shu Shen gratefully acknowledges research support from the Institute for Social Sciences of the University of California, Davis and the Hellman Fellows Award. All errors are the authors'.

Abstract

Treatment effect heterogeneity is frequently studied in regression discontinuity (RD) applications. This paper proposes, under the RD setup, formal tests for treatment effect heterogeneity among individuals with different observed pre-treatment characteristics. The proposed tests study whether a policy treatment 1) is beneficial for at least some subpopulations defined by pre-treatment covariate values, 2) has any impact on at least some subpopulations, and 3) has a heterogeneous impact across subpopulations. The empirical section applies the tests to study the impact of attending a better high school and discovers interesting patterns of treatment effect heterogeneity neglected by previous studies.

JEL classification: C21, C31

Keywords: Sharp regression discontinuity, fuzzy regression discontinuity, treatment effect heterogeneity.

1 Introduction

Regression discontinuity (RD) has gained increasing popularity in the field of applied economics over the past two decades for providing credible and straightforward identification of the causal effect of policies.[1] The identification strategy uses the fact that the probability of an individual receiving a policy treatment changes discontinuously with an underlying variable, often referred to as the running variable. Researchers compare values of the response outcome above and below the point of the underlying variable where the discontinuity occurs and identify the average treatment effect of individuals at the margin of policy treatment. If researchers are only interested in the average effect, controlling for additional covariates other than the running variable is not necessary.[2] However, when researchers are further interested in treatment effect heterogeneity, or how policy impact varies among individuals, it is important to extract information from additional controls and to consider the conditional average policy effects for individuals with different observed characteristics.

This paper considers the inference of conditional average policy effects under the RD setup. Specifically, we propose (uniform) tests that examine whether a policy treatment 1) is beneficial for at least some subpopulations defined by observed pre-treatment covariate values, 2) has any impact on at least some subpopulations, and 3) has a heterogeneous impact across all subpopulations. Both sharp and fuzzy RD designs are considered.

The proposed tests are useful because applied researchers are often interested in treatment effect heterogeneity. A survey of recent publications in top general interest journals in economics finds that 15 out of 17 papers that adopted the RD framework analyzed treatment effect heterogeneity.[3] The common practice is to build a linear regression

[1] Pioneering work and early applications in RD include van der Klaauw (2002), Angrist and Lavy (1999), and Black (1999), among others. Imbens and Lemieux (2008) and Lee and Lemieux (2010) provide excellent reviews of the topic.

[2] Although controlling for additional covariates can potentially improve the estimation efficiency of the RD estimators. See, for example, Imbens and Lemieux (2008) and Calonico et al. (2016).

[3] The survey includes all 2015 issues of the Quarterly Journal of Economics (QJE), Journal of Political Economy (JPE), the Review of Economic Studies (REStud), the American Economic Review (AER), American Economic Journal: Applied Economics (AEJ: AE) and American Economic Journal: Economic Policy (AEJ: EP) as well as 2016 issues of these journals published before April. A total of 17 papers (0 in

model with interaction terms between the discontinuity dummy and additional controls of interest or to accompany the primary RD regression with subsample regressions.

The interaction term method, which adds interaction terms between the dummy variable indicating whether an individual passes the cut-off value of the running variable and additional covariates of interest to the RD regression model, is parametric. The method severely over-rejects under model misspecification even if researchers only use observations close to the cut-off of the running variable for estimation. This is in sharp contrast to the classic RD regression method which is nonparametric and robust to misspecification under certain mild kernel, bandwidth, and smoothness conditions of the underlying distribution.

The subsample regression method repeats the main RD analysis with different subsamples defined by individual observed characteristics. This method is nonparametric. However, as the method typically involves running a number of subsample RD regressions at the same time, it is essential to adjust the regressions for multiple testing (see, e.g., Romano and Shaikh, 2010 and Anderson, 2008) to achieve correct inference. Unfortunately, none of the papers in our survey using the subsample regression method address this issue. Furthermore, even if multiple testing is correctly accounted for, the subsample RD regression method is not ideal. First, under the fuzzy RD design, it can produce over-rejected tests and under-covered confidence intervals if the sample size or proportion of compliers is small for some subsamples. This is because the method uses subsample local average treatment effect estimators, which could have non-classical inference when the first stage is weak (Feir et al., 2015) or when the subsample size is small. Second, to implement the subsample regression method, researchers often categorize continuous covariates into discrete groups in an arbitrary way, which often results in loss of information.

The tests we propose are nonparametric and robust to weak identification. We formulate our null hypotheses as conditional moment equalities/inequalities conditional on

QJE, 1 in JPE, 0 in REStud, 5 in AER, 5 in AEJ: AE, and 6 in AEJ: EP) use the RD method, among which 15 address the issue of treatment effect heterogeneity. 2 of the 15 papers carry out the heterogeneity analysis using linear regressions with interaction terms. All the other 13 papers use subsample RD regressions. None of the 13 papers using the subsample regression method correct for multiple testing.

both the running variable of the RD model and additional controls of interest and then apply the instrument function approach developed in Andrews and Shi (2013, 2014) to transform the hypotheses to conditional moment equalities/inequalities conditional only on the running variable. This transformation is without loss of information, and all transformed moments could be estimated by nonparametric local linear regression at the boundary. The tests have statistics of order $(nh)^{-1/2}$, which means that although we are looking at average policy effects conditional on multiple control variables the statistics have the same rate of convergence as the classic RD estimators that do not control for covariates other than the running variable. Furthermore, the proposed test statistics for the fuzzy RD set-up do not rely on plug-in conditional local average treatment effect estimators, which could have poor small sample performance with conventional asymptotic approximations due to the random denominator problem. We demonstrate with Monte Carlo simulations that the proposed tests have very good small sample performance. In contrast, the interaction term and the subsample RD regression methods currently adopted in the applied literature have very unsatisfactory simulation performance.

We apply the proposed tests to study the impact of attending a better high school in Romania following Pop-Eleches and Urquiola (2013). The mean RD analysis in Pop-Eleches and Urquiola (2013) finds that going to a more selective high school significantly improves the average Baccalaureate exam grade among marginal students but does not seem to affect the probability of a student taking the Baccalaureate exam. Pop-Eleches and Urquiola (2013) carry out a heterogeneity analysis using the subsample regression method and find little evidence of treatment effect heterogeneity. In contrast, our tests detect a clear signal of treatment effect heterogeneity. We find that attending a more selective high school has a positive impact on the exam-taking rate for some subpopulations and a negative impact for some other subpopulations. Our results suggest that the insignificant mean effect on the exam-taking rate documented in Pop-Eleches and Urquiola (2013) may come from the cancellation of opposite-signed effects among different schools.

This paper is related to the large literature on regression discontinuity, especially some recent developments that also look at treatment effect heterogeneity. For example, Frandsena et al. (2012) and Shen and Zhang (2016) study treatment effect heterogeneity associated with unobserved individual characteristics and the distributional treatment ef-

fect under the RD framework; Dong and Lewbel (2015) and Angrist and Rokkanen (2015) study treatment effect extrapolation away from the running variable cut-off; Bertanha (2016) and Cattaneo et al. (2016) study treatment effect heterogeneity at different values of the running variable when the RD design has multiple cut-offs; and Bertanha and Imbens (2014) examine the external validity of the local average treatment effect under fuzzy RD by testing whether treated/untreated compliers and always-takers/never-takers have equal distributions of potential outcomes. Although our paper also studies treatment effect heterogeneity under the RD framework, our focus is different from the work mentioned above. Specifically, we examine heterogeneity of the treatment effect as a function of individual characteristics while other papers look at heterogeneity as a function of the running variable, compliance type, or rank in the potential outcome distributions.

The tests proposed in this paper are related to Andrews and Shi (2013, 2014) and other conditional moment equality/inequality tests that apply the instrument function method. Since estimation of the nonparametric RD model involves boundary estimators in the local polynomial class, and such estimators have not been previously used in conjunction with the instrument function method, our paper contributes to the literature in developing new testing procedures for conditional moment equality/inequalities that require nonparametric boundary estimation. In addition, we propose a new multiplier bootstrap method for simulating critical values in such tests. For the uniform sign test, we discuss both critical values based on the least favorable configuration (LFC) and the generalized moment selection (GMS) method introduced by Andrews and Soares (2010) and Andrews and Shi (2013, 2014, 2017). The GMS method is similar to the recentering method in Hansen (2005) and Donald and Hsu (2016) and the contact set approach proposed in Linton et al. (2010), Lee et al. (2013), and Aradillas-Lopez et al. (2016).

The paper is organized as follows. Section 2 sets up the model and discusses the identification results of the conditional treatment effects of interest under both sharp and fuzzy RD designs. Section 3 proposes three uniform tests for treatment effect heterogeneity under the sharp RD design. Section 4 extends the tests to the fuzzy RD design. Section 5 examines the small sample performance of the proposed tests and compares the performance with other tests currently adopted in the applied literature. Section 6 applies the proposed tests to study the heterogeneous effect of going to a better school.

Proofs are provided in the online appendix.

2 Model Framework

Let Y_i denote the outcome of interest and T_i the treatment decision of individual i. Use $Y_i(0)$ and $Y_i(1)$ to denote potential outcomes when $T_i = 0$ and $T_i = 1$, respectively; $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$. Let Z_i be the running variable and X_i the set of pre-treatment covariates with compact support $\mathcal{X} \subset R^{d_x}$. Without loss of generality, assume that $\mathcal{X} = \times_{j=1}^{d_x}[0,1]$ and use $\mathcal{X}_c \subset \mathcal{X}$ to denote the support of X_i conditional on $Z_i = c$. For notational simplicity, we assume that X_i includes only continuous variables. At the end of Section 3, we will discuss how to implement the proposed tests when X_i contains discrete variables. For any $\delta > 0$, let $\mathcal{N}_{\delta,z}(c) = \{z : \ |z - c| \leq \delta\}$ denote a neighborhood of Z_i around $Z_i = c$.

Assumption 2.1 *The running variable Z_i is continuously distributed in a neighborhood of the threshold value c, where c is an interior point of its support. Also, assume that for some $\delta > 0$,*

(i) *$E[Y_i(1)|X_i = x, Z_i = z]$ and $E[Y_i(0)|X_i = x, Z_i = z]$ are continuous in x and z on $\mathcal{X}_c \times \mathcal{N}_{\delta,z}(c)$;*

(ii) *$f_{x|z}(x|z)$, the density function of $X_i|Z_i = z$, is continuous in z and x on $\mathcal{N}_{\delta,z}(c) \times \mathcal{X}_c$ and is uniformly bounded.*

Assumption 2.1.(i) requires that the means of the potential outcomes conditional on both the running variable and the additional controls of interest are continuous. It is stronger than the standard continuity assumption of $E[Y_i(t)|Z_i = z]$ used in the literature (c.f. Imbens and Lemieux, 2008) for the identification of the average treatment effect, or ATE. Assumption 2.1.(i) is required to identify the conditional average treatment effect, or CATE, as a function of the values of X_i. Assumption 2.1.(ii) requires that the conditional distribution of the additional controls conditional on the running variable is continuous. Although the condition is not required in the literature for the identification of the ATE, it is, in fact, a direct implication of the "no precise control over the running

variable" rule introduced by Lee and Lemieux (2010) and well-accepted in the applied RD literature.[4] Assumption 2.1.(ii) is essential to our testing method where we transform our null hypotheses that condition on both X_i and Z_i to moment equalities/inequalities that condition only on Z_i.

When the treatment decision T_i is a deterministic function of the running variable Z_i such that $T_i = 1(Z_i \geq c)$, the model follows a *sharp RD design*. Under Assumption 2.1, the ATE conditional on $Z_i = c$ and the CATE conditional on both $Z_i = c$ and $X_i = x$ are identified, respectively, as

$$\begin{aligned} ATE &= E[Y_i(1) - Y_i(0)|Z_i = c] = \lim_{z \searrow c} E[Y_i|Z_i = z] - \lim_{z \nearrow c} E[Y_i|Z_i = z], \\ CATE(x) &= E[Y_i(1) - Y_i(0)|X_i = x, Z_i = c] \\ &= \lim_{z \searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i|X_i = x, Z_i = z]. \end{aligned} \tag{2.1}$$

Proofs for identification are given in the online appendix. More generally, when the treatment status T_i is a probabilistic function of Z_i, the RD model follows a *fuzzy* design. Suppose a policy intervention encourages an individual i to receive the treatment if the running variable Z_i is greater than or equal to c. Let $T_i(1)$ and $T_i(0)$ be the potential treatment decisions of individual i depending on whether he/she is encouraged or not, and $T_i = T_i(1)1(Z_i \geq c) + T_i(0)1(Z_i < c)$. For identification in this general case we require the following assumptions in replacement of Assumption 2.1.

Assumption 2.2 *The running variable Z_i is continuously distributed in a neighborhood around the threshold value c, where c is an interior point of its support. Also, assume that for some $\delta > 0$,*

(i) *$E[Y_i(t)|T_i(1) - T_i(0) = 1, X_i = x, Z_i = z]$ and $E[Y_i(t)|T_i(1) = T_i(0) = t', X_i = x, Z_i = z]$ are continuous in x and z on $\mathcal{X}_c \times \mathcal{N}_{\delta,z}(c)$ for $t, t' \in \{0, 1\}$;*

(ii) *$P[T_i(1) - T_i(0) = 1|X_i = x, Z_i = z]$ and $P[T_i(1) = T_i(0) = t|X_i = x, Z_i = z]$ are continuous in x and z on $\mathcal{X}_c \times \mathcal{N}_{\delta,z}(c)$ for $t \in \{0, 1\}$;*

[4] According to Lee and Lemieux (2010), an individual is said to have imprecise control over the running variable if the conditional density $Z_i = z|(X_i, V_i)$ is continuous in z around c, where V_i represents unobserved characteristics of individual i. By Bayes' Rule, this condition implies that the density of $(X_i, V_i)|Z_i = z$ is continuous in z around c, which further implies continuity of the density $X_i|Z_i = z$ around $z = c$.

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

(iv) $E[T_i(1) - T_i(0)|X_i = x, Z_i = c] > 0$ *for all* $x \in \mathcal{X}_c$*;*

(v) $f_{x|z}(x|z)$*, the density function of* $X_i|Z_i = z$*, is continuous in* z *and* x *on* $\mathcal{N}_{\delta,z}(c) \times \mathcal{X}_c$ *and is uniformly bounded above.*

Assumption 2.2.(i) requires the continuity of average potential outcomes for always-takers, compliers, and never-takers with respect to both the running variable Z_i and the additional control X_i. Assumption 2.2.(ii) requires the continuity of the probability of an individual belonging to each group. Assumption 2.2.(iii) and (iv) require no defiers and a non-trivial presence of compliers, respectively. Assumption 2.2.(i), (ii) and (iv) are stronger than their counterparts that are unconditional on X_i (c.f. Dong and Lewbel, 2015). Assumption 2.2.(iii) is the monotonicity restriction that is commonly required in fuzzy RD models. It implies that $E[T_i(1) - T_i(0)|X_i = x, Z_i = c] \geq 0$ for all $x \in \mathcal{X}_c$, or that the first-stage effect is positive for all observationally equivalent individuals at the treatment cut-off. It is worth noting that this testable implication of the monotonicity assumption could be tested by one of the tests proposed in the next section. Assumption 2.2.(v) is the same as Assumption 2.1.(ii).

Under the fuzzy RD design, the local average treatment effect, or LATE, and the conditional local average treatment effect for compliers, or CLATE, are identified as

$$\begin{aligned} LATE =& E[Y_i(1) - Y_i(0)|Z_i = c, T_i(1) - T_i(0) = 1] \\ =& \frac{\lim_{z \searrow c} E[Y_i|Z_i = z] - \lim_{z \nearrow c} E[Y_i|Z_i = z]}{\lim_{z \searrow c} E[T_i|Z_i = z] - \lim_{z \nearrow c} E[T_i|Z_i = z]}, \\ CLATE(x) =& E[Y_i(1) - Y_i(0)|X_i = x, Z_i = c, T_i(1) - T_i(0) = 1] \\ =& \frac{\lim_{z \searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i|X_i = x, Z_i = z]}{\lim_{z \searrow c} E[T_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[T_i|X_i = x, Z_i = z]}. \end{aligned} \tag{2.2}$$

Proofs for identification are again given in the online appendix. All identified treatment effects, including the ATE, $LATE$, $CATE$, and $CLATE$, can be estimated by standard local linear estimation methods.

3 Testing Under the Sharp RD Design

Researchers are often interested in knowing 1) whether a policy treatment is beneficial for at least some subpopulations defined by pre-treatment covariate values, 2) whether the policy treatment has any impact on at least some subpopulations, and 3) whether the policy's effect is heterogeneous across all subpopulations. In this section, we develop uniform tests for these purposes under the sharp RD design. We extend the tests to the fuzzy RD case in the next section.

3.1 Testing if the Treatment is Beneficial for At Least Some Subpopulations

Hypotheses Formation

To test if a policy treatment is on average beneficial to at least some subpopulations defined by covariate values, the null and alternative hypotheses can be formulated as

$$\begin{aligned} &H_{0,ate}^{neg} : CATE(x) = E[Y_i(1) - Y_i(0)|X_i = x, Z_i = c] \leq 0, \ \forall \ x \in \mathcal{X}_c, \\ &H_{1,ate}^{neg} : CATE(x) = E[Y_i(1) - Y_i(0)|X_i = x, Z_i = c] > 0, \text{ for some } \ x \in \mathcal{X}_c. \end{aligned} \tag{3.1}$$

The null and alternative hypotheses $H_{0,ate}^{neg}$ and $H_{1,ate}^{neg}$ are defined by conditional moment inequalities conditional on both the running variable Z and the additional control X. We apply the instrument function approach in Andrews and Shi (2013, 2014) to transform these inequalities to an infinite number of instrumented moment inequalities conditional on only the running variable Z, without loss of information.

Let $\mathcal{G}$ be the set of the indicator functions of countable hypercubes C_ℓ such that

$$\begin{aligned} &\mathcal{G} = \{g_\ell(\cdot) = 1(\cdot \in C_\ell) : \ell \equiv (x, r) \in \mathcal{L}\}, \text{ where} \\ &C_\ell = \left(\times_{j=1}^{d_x}(x_j, x_j + r]\right) \cap \mathcal{X} \text{ and} \\ &\mathcal{L} = \left\{(x, q^{-1}) : \ q \cdot x \in \{0, 1, 2, \cdots, q-1\}^{d_x}, \text{ and } q = 1, 2, \cdots\right\}. \end{aligned} \tag{3.2}$$

For each $\ell \in \mathcal{L}$, define the instrumented moment condition $\nu(\ell)$ by

$$\nu(\ell) = E[g_\ell(X_i)CATE(X_i)|Z_i = z],$$

which represents the average treatment effect for individuals with $X_i \in C_\ell$ multiplied by those individuals' proportion of the total population. The following lemma shows that hypotheses $H^{neg}_{0,ate}$ and $H^{neg}_{1,ate}$ can be characterized by the following instrumented conditional moment inequalities without loss of information.

$$\begin{aligned} &H^{neg}_{0,ate}: \ \nu(\ell) \leq 0, \ \forall\, \ell \in \mathcal{L}, \\ &H^{neg}_{1,ate}: \ \nu(\ell) > 0, \ \text{for some } \ell \in \mathcal{L}. \end{aligned} \tag{3.3}$$

Lemma 3.1 *Under Assumption 2.1, the hypotheses in (3.1) are equivalent to those in (3.3).*

Notice that when $q = 1$, $\ell = (\mathbf{0}, 1)$, $C_\ell = (0,1]^{d_x} = \mathcal{X}$, $\nu(\ell)$ reduces to $\nu((\mathbf{0},1)) = E[CATE(X_i)|Z_i = z] = ATE$. When $q = 2$, the side length of the hypercubes is $1/2$. Suppose that $d_x = 2$, then there are four possible values of ℓ's: $((0,0), 1/2)$, $((1/2,1/2), 1/2)$, $((0,1/2), 1/2)$ and $((1/2,0), 1/2)$. They correspond to four hypercubes or C_ℓ's: $(0,1/2]^2$, $(1/2,1]^2$, $(0,1/2] \times (1/2,1]$ and $(1/2,1] \times (0,1/2]$. When q gets larger, the hypercubes get smaller.

Test Statistic and Asymptotic Results

Under Assumption 2.1 and by standard RD identification strategy, we know that $\nu(\ell)$ is identified by

$$\nu(\ell) = \lim_{z \searrow c} E[g_\ell(X_i)Y_i|Z_i = z] - \lim_{z \nearrow c} E[g_\ell(X_i)Y_i|Z_i = z], \tag{3.4}$$

for each $\ell \in \mathcal{L}$. Then, by standard RD estimation strategy, $\nu(\ell)$ could be estimated by

$$\hat{\nu}(\ell) = \hat{m}_+(\ell) - \hat{m}_-(\ell),$$

where $\hat{m}_+(\ell)$ and $\hat{m}_-(\ell)$ are local linear estimators of $m_+(\ell) = \lim_{z \searrow c} E[g_\ell(X_i)Y_i|Z_i = z]$ and $m_-(\ell) = \lim_{z \nearrow c} E[g_\ell(X_i)Y_i|Z_i = z]$. Let $K(\cdot)$ be the kernel function and h the bandwidth. The estimators $\hat{m}_+(\ell)$ and $\hat{m}_-(\ell)$ could be defined as

$$(\hat{m}_+(\ell), \hat{b}_+(\ell)) = \arg\min_{a,b} \sum_{i=1}^{n} 1(Z_i \geq c) \cdot K\left(\frac{Z_i - c}{h}\right) \Big[g_\ell(X_i)Y_i - a - b \cdot (Z_i - c)\Big]^2,$$

$$(\hat{m}_-(\ell), \hat{b}_-(\ell)) = \arg\min_{a,b} \sum_{i=1}^{n} 1(Z_i < c) \cdot K\left(\frac{Z_i - c}{h}\right) \Big[g_\ell(X_i)Y_i - a - b \cdot (Z_i - c)\Big]^2.$$

Following Fan and Gijbels (1992), for $j = 0, 1, 2, \ldots$, define

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

and re-write the local linear estimators as

$$\hat{m}_{+}(\ell) = \frac{\sum_{i=1}^{n} 1(Z_i \geq c) K(\frac{Z_i - c}{h})[S_{n,2}^{+} - S_{n,1}^{+}(Z_i - c)] g_\ell(X_i) Y_i}{S_{n,0}^{+} S_{n,2}^{+} - S_{n,1}^{+} S_{n,1}^{+}} \equiv \sum_{i=1}^{n} w_{ni}^{+} \cdot g_\ell(X_i) Y_i,$$

$$\hat{m}_{-}(\ell) = \frac{\sum_{i=1}^{n} 1(Z_i < c) K(\frac{Z_i - c}{h})[S_{n,2}^{-} - S_{n,1}^{-}(Z_i - c)] g_\ell(X_i) Y_i}{S_{n,0}^{-} S_{n,2}^{-} - S_{n,1}^{-} S_{n,1}^{-}} \equiv \sum_{i=1}^{n} w_{ni}^{-} \cdot g_\ell(X_i) Y_i,$$

where $w_{ni}^{+} = \frac{1(Z_i \geq c) K(\frac{Z_i - c}{h})[S_{n,2}^{+} - S_{n,1}^{+}(Z_i - c)]}{S_{n,0}^{+} S_{n,2}^{+} - S_{n,1}^{+} S_{n,1}^{+}}$ and $w_{ni}^{-} = \frac{1(Z_i < c) K(\frac{Z_i - c}{h})[S_{n,2}^{-} - S_{n,1}^{-}(Z_i - c)]}{S_{n,0}^{-} S_{n,2}^{-} - S_{n,1}^{-} S_{n,1}^{-}}$.

Next we discuss the asymptotic properties of above described nonparametric estimators. Let f_z be the density function of Z_i and $f_{xz}(x, z)$ the joint density of X_i and Z_i. Let $\vartheta_j = \int_0^\infty u^j K(u) du$ for $j = 0, 1, 2, \ldots$, $\sigma_{+}^2(\ell_1, \ell_2) = \lim_{z \searrow c} Cov[g_{\ell_1}(X_i) Y_i, g_{\ell_2}(X_i) Y_i | Z_i = z]$, and $\sigma_{-}^2(\ell_1, \ell_2) = \lim_{z \nearrow c} Cov[g_{\ell_1}(X_i) Y_i, g_{\ell_2}(X_i) Y_i | Z_i = z]$. Let $\mu_d(x, z) = E[Y_i(d) | X_i = x, Z_i = z]$, $\sigma_d^2(x, z) = Var(Y_i(d) | X_i = x, Z_i = z)$ for $d = 0, 1$, and $\mathcal{X}_z$ be the support of X conditioning on $Z_i = z$. We make the following assumptions.

Assumption 3.1 *Assume that there exists $\delta > 0$ such that (i) $\mathcal{X}_z = \mathcal{X}_c$ for all $z \in \mathcal{N}_{\delta,z}(c)$; (ii) $f_z(z)$ is twice continuously differentiable in z on $\mathcal{N}_{\delta,z}(c)$; (iii) $f_z(z)$ is bounded away from zero on $\mathcal{N}_{\delta,z}(c)$; (iv) for each $x \in \mathcal{X}_c$, $f_{xz}(x, z)$ is twice continuously differentiable in z on $\mathcal{N}_{\delta,z}(c)$; (v) $|\partial^2 f_{xz}(x, z)/\partial z \partial z|$ is uniformly bounded on $x \in \mathcal{X}_c$ and $z \in \mathcal{N}_{\delta,z}(c)$; (vi) for $d = 0, 1$, and for each $x \in \mathcal{X}_c$, $\mu_d(x, z)$ is twice continuously differentiable in z on $\mathcal{N}_{\delta,z}(c)$; (vii) for $d = 0, 1$, $|\partial^2 \mu_d(x, z)/\partial z \partial z|$ is uniformly bounded on $x \in \mathcal{X}_c$ and $z \in \mathcal{N}_{\delta,z}(c)$; (viii) $E[Y_i^4 | Z_i = z]$ is uniformly bounded for z on $\mathcal{N}_{\delta,z}(c)$; (ix) for $d = 0, 1$, $\sigma_d^2(x, z)$ is uniformly bounded on $x \in \mathcal{X}_c$ and $z \in \mathcal{N}_{\delta,z}(c)$.*

Assumption 3.2 *Assume that (i) the function $K(\cdot)$ is a non-negative symmetric bounded kernel with a compact support; (ii) $\int K(u) du = 1$; (iii) $h \to 0$, $nh \to \infty$ and $nh^5 \to 0$ as $n \to \infty$.*

Assumption 3.3 *Let $\{U_i : 1 \leq i \leq n\}$ be a sequence of i.i.d. random variables where $E[U_i] = 0$, $E[U_i^2] = 1$, and $E[U_i^4] < M$ for some $\delta > 0$ and $M > 0$. $\{U_i : 1 \leq i \leq n\}$ is independent of the sample path $\{(Y_i, X_i, Z_i, T_i) : 1 \leq i \leq n\}$.*

Assumption 3.1 states smoothness conditions of the underlying data distribution. Assumption 3.1(i) is assumed for notational simplicity. We can allow $\mathcal{X}_z$ to depend on z and all the proofs will still go through with much more involved notations. Assumption 3.1(ii)-(vi) are standard smoothness conditions for local linear estimation. Assumption 3.1(vii) regulates the bias of $\hat{\nu}(\ell)$ to be uniformly asymptotically negligible. Assumptions 3.1($viii$) and (ix) regulate the estimator of the covariance kernel of the limiting process to be uniformly consistent. Similar conditions are used in Andrews and Shi (2014) and Hsu (2017). Assumption 3.2(i) and (ii) are standard conditions on the kernel function. The triangular kernel which is most frequently adopted in RD regressions satisfies these conditions. Assumption 3.2(iii) requires undersmoothed bandwidth so that $\sqrt{nh}\,(\hat{\nu}(\cdot) - \nu(\cdot))$ weakly converges to a mean zero Gaussian process. Assumption 3.3 is required for the validity of the multiplier bootstrap.

Given the assumptions, we can summarize the asymptotic properties of $\hat{\nu}(\cdot)$ in the following lemma.

Lemma 3.2 *Under Assumption 2.1, and Assumptions 3.1-3.2, we have*

$$\left|\sqrt{nh}\,(\hat{\nu}(\ell) - \nu(\ell)) - \sum_{i=1}^{n} \phi_{\nu,ni}(\ell)\right| = o_p(1),$$

$$\phi_{\nu,ni}(\ell) = \sqrt{nh}\left(w_{ni}^{+} \cdot (g_\ell(X_i)Y_i - m_+(\ell)) - w_{ni}^{-} \cdot (g_\ell(X_i)Y_i - m_-(\ell))\right), \tag{3.5}$$

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

$$\widehat{\Phi}_{\nu,n}(\cdot) \equiv \sqrt{nh}(\hat{\nu}(\cdot) - \nu(\cdot)) \Rightarrow \Phi_{h_{2,\nu}}(\cdot),$$

where $\Phi_{h_2,\nu}(\cdot)$ denotes a mean zero Gaussian process with covariance kernel

$$h_{2,\nu}(\ell_1, \ell_2) = \frac{\int_0^\infty (\vartheta_2 - u\vartheta_1)^2 K^2(u)du}{(\vartheta_2\vartheta_0 - \vartheta_1^2)^2} \frac{\sigma_+^2(\ell_1, \ell_2) + \sigma_-^2(\ell_1, \ell_2)}{f_z(c)},$$

for $\ell_1, \ell_2 \in \mathcal{L}$.

The $\phi_{\nu,ni}(\ell)$ function defined in (3.5) is the influence function used to derive the limiting distribution of $\sqrt{nh}\,(\hat{\nu}(\ell) - \nu(\ell))$. Let $\hat{\sigma}_{\nu,n}^2(\ell) = \sum_{i=1}^n \hat{\phi}_{\nu,ni}(\ell)^2$ where $\hat{\phi}_{\nu,ni}(\ell) = \sqrt{nh}\left(w_{ni}^{+} \cdot (g_\ell(X_i)Y_i - \hat{m}_+(\ell)) - w_{ni}^{-} \cdot (g_\ell(X_i)Y_i - \hat{m}_-(\ell))\right)$ is the estimated influence function that replaces $m_+(\ell)$ and $m_-(\ell)$ in $\phi_{\nu,ni}(\ell)$ by their nonparametric estimators. We will show that $\hat{\sigma}_{\nu,n}^2(\ell)$ is a consistent estimator for $\sigma_\nu^2(\ell) \equiv h_{2,\nu}(\ell, \ell)$ uniformly over $\ell \in \mathcal{L}$.

Define $\hat{\sigma}^2_{\nu,\epsilon}(\ell) = \max\{\hat{\sigma}^2_{\nu,n}(\ell), \epsilon \cdot \hat{\sigma}^2_{\nu,n}(\mathbf{0}, 1)\}$ with some small positive ϵ that constrains the variance estimator to be positive. We define the scale-invariant Kolmogorov-Smirnov (KS) type statistic for testing $H^{neg}_{0,ate}$ as

$$\widehat{S}^{neg}_{ate} = \sqrt{nh} \sup_{\ell \in \mathcal{L}} \frac{\hat{\nu}_n(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)}.$$

In the simulation and empirical sections of the paper we follow the practice in Andrews and Shi (2013) and set ϵ to 0.05 .

Notice that we adopted the instrument function approach in Andrews and Shi (2013, 2014) to test the conditional moment inequalities stated in $H^{neg}_{0,ate}$. Other methods developed in the conditional moment inequality literature (e.g., Chernozhukov et al., 2013; Lee et al., 2013, 2017; Aradillas-Lopez et al., 2016; Chetverikov, 2018) could also be potentially applied to test the null. The simulation results in Aradillas-Lopez et al. (2016) suggest that no specific strategy is expected to outperform the rest in all data generating processes. We choose to use the instrument function approach because it transforms the null to a series of conditional moment inequalities that only condition on the running variable Z. The constructed test statistic then only involves one-dimensional local linear estimation, which is the same as the estimation strategy adopted for classic RD regressions.

Simulated Critical Value Based on the Least Favorable Configuration

Given the influence function representation in (3.5), we can use a multiplier bootstrap method (see, e.g., Hsu, 2016) to approximate the whole empirical process. To be specific, let U_1, U_2,... be i.i.d. pseudo random variables with $E[U] = 0$, $E[U^2] = 1$, and $E[U^4] < \infty$ that are independent of the sample path. Let the simulated process $\widehat{\Phi}^u_{\nu,n}(\ell)$ be

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

The next lemma shows that the process $\widehat{\Phi}^u_{\nu,n}(\cdot)$ can approximate the empirical process $\widehat{\Phi}_{\nu,n}(\cdot)$ well. The proofs are given in the online appendix.

Lemma 3.3 *Under Assumption 2.1 and Assumptions 3.1-3.3,* $\sup_{\ell \in \mathcal{L}} |\hat{\sigma}^2_{\nu,n}(\ell) - \sigma^2_{\nu}(\ell)| \xrightarrow{p} 0$ *and* $\widehat{\Phi}^u_n(\cdot) \overset{p}{\Rightarrow} \Phi_{h_{2,\nu}}(\cdot)$.[5]

[5]The conditional weak convergence is in the sense of Section 2.9 of van der Vaart and Wellner (1996)

In the simulation and empirical sections of the paper, the pseudo random variables are drawn from the standard normal distribution. Let P^u denote the multiplier probability measure. For significance level $\alpha < 1/2$, define the simulated critical value $\hat{c}^{neg}_{n,ate}(\alpha)$ as

$$\hat{c}^{neg}_{n,ate}(\alpha) = \sup\left\{ q \Big| P^u\Big(\sup_{\ell\in\mathcal{L}} \frac{\widehat{\Phi}^u_{\nu,n}(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)} \le q \Big) \le 1-\alpha \right\},$$

i.e., $\hat{c}^{neg}_{n,ate}(\alpha)$ is the $(1-\alpha)$-th quantile of the simulated null distribution, $\sup_{\ell\in\mathcal{L}} \frac{\widehat{\Phi}^u_{\nu,n}(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)}$. The *decision rule* of the test is then defined as: "Reject $H^{neg}_{0,ate}$ if $\widehat{S}^{neg}_{ate} > \hat{c}^{neg}_{n,ate}(\alpha)$."

Generalized Moment Selection

The above-described testing procedure relies on the least favorable configuration, or LFC, and could be potentially conservative. In this section, we follow Andrews and Shi (2013) and apply the GMS method to improve the power of the proposed test.

Assumption 3.4 *Let a_n and B_n be sequences of non-negative numbers.*

1. *a_n satisfies that $\lim_{n\to\infty} a_n = \infty$, and $\lim_{n\to\infty} a_n/\sqrt{nh} = 0$.*
2. *B_n is non-decreasing and satisfies that $\lim_{n\to\infty} B_n = \infty$, and $\lim_{n\to\infty} B_n/a_n = 0$.*

Let η be a small positive number. For all $\ell \in \mathcal{L}$, define

$$\hat{\psi}_\nu(\ell) = -B_n \cdot 1\left(\sqrt{nh} \cdot \frac{\hat{\nu}_n(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)} < -a_n \right), \tag{3.6}$$

and the simulated GMS critical value $\hat{c}^{\eta}_{n,ate}(\alpha)$ as

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

Then $\hat{c}^{\eta}_{n,ate}(\alpha)$ is the $(1-\alpha+\eta)$-th quantile of the supremum $\left(\frac{\widehat{\Phi}^u_{\nu,n}(\ell)}{\hat{\sigma}_{\mu,\epsilon}(\ell)} + \hat{\psi}_\nu(\ell) \right)$, plus η. In the simulation and empirical sections of the paper, we follow Andrews and Shi (2013, 2014) and use $a_n = (0.3\ln(n))^{1/2}$, $B_n = (0.4\ln(n)/\ln\ln(n))^{1/2}$ and $\eta = 10^{-6}$.

and Chapter 2 of Kosorok (2008). To be more specific, $\Psi^u_n \overset{p}{\Rightarrow} \Psi$ in the metric space $(\mathbb{D}, d)$ if and only if $\sup_{f\in BL_1} |E_u f(\Psi^u_n) - Ef(\Psi)| \xrightarrow{p} 0$ and $E_u f(\Psi^u_n)^* - E_u f(\Psi^u_n)_* \xrightarrow{p} 0$, where the subscript u in E_u indicates conditional expectation over the U_i's given the remaining data, BL_1 is the space of functions $f : \mathbb{D} \to R$ with Lipschitz norm bounded by 1, and $f(\Psi^u_n)^*$ and $f(\Psi^u_n)_*$ denote measurable majorants and minorants with respect to the joint data including the U_i's.

Let the *decision rule* based on the GMS critical value be: "Reject $H_{0,ate}^{neg}$ if $\widehat{S}_{ate}^{neg} > \hat{c}_{n,ate}^{\eta}(\alpha)$." Intuitively, the term $\hat{\psi}_{\nu}(\ell)$ helps to suppress the influence of negative moment functions on the simulated critical value as the term is negative for ℓ vectors with large and negative $\hat{\nu}_n(\ell)/\hat{\sigma}_{\nu,\epsilon}(\ell)$ values and is zero otherwise. Using the GMS critical value hence improves the power of the proposed inequality test.

Size and Power Properties

We summarize size and power properties of the proposed test in the following two theorems. The proofs are given in the appendix.

Theorem 3.1 *Under Assumption 2.1 and Assumptions 3.1-3.3, when $\alpha < 1/2$, we have*

(1) under $H_{0,ate}^{neg}$, $\lim_{n\to\infty} P(\widehat{S}_{ate}^{neg} > \hat{c}_{n,ate}^{neg}(\alpha)) \leq \alpha$, and

(2) under $H_{1,ate}^{neg}$, $\lim_{n\to\infty} P(\widehat{S}_{ate}^{neg} > \hat{c}_{n,ate}^{neg}(\alpha)) = 1$.

Theorem 3.1 discusses the asymptotic properties of the proposed test based on the LFC critical value. The test is consistent and its asymptotic size is less than or equal to the significance level α as a result of adopting the LFC critical value.

Theorem 3.2 *Under Assumption 2.1 and Assumptions 3.1-3.4, when $\alpha < 1/2$, we have*

(1a) under $H_{0,ate}^{neg}$, $\lim_{n\to\infty} P(\widehat{S}_{n,ate}^{neg} > \hat{c}_{n,ate}^{\eta}(\alpha)) \leq \alpha$,

(1b) if $\mathcal{L}^o = \{\ell : \nu(\ell) = 0\}$ is non-empty and there exists $\ell^* \in \mathcal{L}^o$ with $\sigma_{\nu}^2(\ell^*) > 0$, then under $H_{0,ate}^{neg}$, $\lim_{\eta\to 0}\lim_{n\to\infty} P(\widehat{S}_{n,ate} > \hat{c}_{n,ate}^{\eta}(\alpha)) = \alpha$, and

(2) under $H_{1,ate}^{neg}$, $\lim_{n\to\infty} P(\widehat{S}_{ate}^{neg} > \hat{c}_{n,ate}^{\eta}(\alpha)) = 1$.

Theorem 3.2 shows the consistency and asymptotic size control of the proposed test based on the GMS critical value. When the null hypothesis $H_{0,ate}^{neg}$ holds with equality for some ℓ vectors, using the GMS critical value can lead to exact asymptotic size control.

In addition, we would like to point out that the above described test for $H_{0,ate}^{neg}$ can be trivially extended to study the hypotheses$H_{0,ate}^{pos} : CATE(x) \geq 0,\ \forall\ x \in \mathcal{X}_c$ in any sharp RD design, or the first stage selection $H_{0,fs}^{pos} : E[T_i(1) - T_i(0)|X_i = x, Z_i = c] \geq 0,\ \forall\ x \in \mathcal{X}_c$

in any fuzzy RD design. As is discussed in the introduction, the second test above is a sufficient test for the monotonicity restriction commonly used in fuzzy RD models in the sense that if $H_{0,fs}^{pos}$ is rejected, the monotonicity assumption is rejected.

Adding Discrete Covariates to the Control Set

Although we have so far restricted the X_i variable to be continuous, the tests we propose can be easily adapted to the case where X_i includes discrete covariates. Without loss of generality, consider the case where in addition to X_i, there is one binary variable, X_{di}, of interest. Let $\mathcal{G}_1 \equiv \{1(X_{di} = 1) \cdot g_\ell(\cdot) : \ \ell \in \mathcal{L}\}$ and $\mathcal{G}_0 \equiv \{1(X_{di} = 0) \cdot g_\ell(\cdot) : \ \ell \in \mathcal{L}\}$. Let $\widetilde{\mathcal{G}} = \mathcal{G} \cup \mathcal{G}_1 \cup \mathcal{G}_0$. It is straightforward to show that hypotheses

$$
\begin{aligned}
&H_{0,ate}^{neg} : CATE(x, x_d) \leq 0, \ \forall \ x \in \mathcal{X}_c \text{ and } x_d \in \{0, 1\}, \\
&H_{1,ate}^{neg} : CATE(x, x_d) > 0, \text{ for some } \ x \in \mathcal{X}_c \text{ and } \ x_d \in \{0, 1\}.
\end{aligned}
$$

are equivalent to

$$
\begin{aligned}
&H_{0,ate}^{neg} : \ \nu(\tilde{g}) = E[\tilde{g}(X_i, X_{di}) CATE(X_i, X_{di})] \leq 0, \ \forall \ \tilde{g} \in \widetilde{\mathcal{G}}, \\
&H_{1,ate}^{neg} : \ \nu(\tilde{g}) = E[\tilde{g}(X_i, X_{di}) CATE(X_i, X_{di})] > 0, \text{ for some } \tilde{g} \in \widetilde{\mathcal{G}}.
\end{aligned}
$$

Then we can carry out the uniform sign test in the same way as is discussed earlier but with $\mathcal{G}$ replaced by $\widetilde{\mathcal{G}}$. All results discussed earlier will remain valid.

3.2 Testing if the Treatment Has Any Impact

To test if a policy treatment has any impact on at least some subpopulations, the null and alternative hypotheses can be formulated as

$$
\begin{aligned}
&H_{0,ate}^{zero} : CATE(x) = 0, \ \forall \ x \in \mathcal{X}_c, \\
&H_{1,ate}^{zero} : CATE(x) \neq 0, \text{ for some } \ x \in \mathcal{X}_c.
\end{aligned} \tag{3.7}
$$

Similar to the previous subsection, we can transform the hypotheses in (3.7) to

$$
\begin{aligned}
&H_{0,ate}^{zero} : \ \nu(\ell) = 0, \ \forall \ \ell \in \mathcal{L}, \\
&H_{1,ate}^{zero} : \ \nu(\ell) \neq 0, \text{ for some } \ell \in \mathcal{L}
\end{aligned} \tag{3.8}
$$

without loss of information, as is summarized in the following lemma.

Lemma 3.4 *Suppose that Assumption 2.1 holds. Then the hypotheses in (3.7) are equivalent to those in (3.8).*

Define the KS type test statistic as

$$\widehat{S}_{ate}^{zero} = \sqrt{nh} \sup_{\ell \in \mathcal{L}} \frac{\left|\hat{\nu}_n(\ell)\right|}{\hat{\sigma}_{\nu,\epsilon}(\ell)}$$

and let the decision rule be: "Reject $H_{0,ate}^{zero}$ if $\widehat{S}_{ate}^{zero} > \hat{c}_{n,ate}^{zero}(\alpha)$", where α is the predetermined significance level and $\hat{c}_{n,ate}^{zero}(\alpha)$ is the simulated critical value defined as

$$\hat{c}_{n,ate}^{zero}(\alpha) = \sup\left\{ q \Big| P^u\Big(\sup_{\ell \in \mathcal{L}} \frac{\left|\widehat{\Phi}_{\nu,n}^u(\ell)\right|}{\hat{\sigma}_{\nu,\epsilon}(\ell)} \leq q \Big) \leq 1 - \alpha \right\}.$$

The following theorem summarizes the size and power properties of the test.

Theorem 3.3 *Under Assumption 2.1 and Assumptions 3.1-3.3, when $\alpha < 1/2$, we have*

(1) under $H_{0,ate}^{zero}$, $\lim_{n\to\infty} P(\widehat{S}_{ate}^{zero} > \hat{c}_{n,ate}^{zero}(\alpha)) = \alpha$, and

(2) under $H_{1,ate}^{zero}$, $\lim_{n\to\infty} P(\widehat{S}_{ate}^{zero} > \hat{c}_{n,ate}^{zero}(\alpha)) = 1$.

Since the null hypothesis $H_{0,ate}^{zero}$ involves only conditional moment equality constraints, the asymptotic convergence results given in Lemmas 3.2 and 3.3 imply that the proposed test for $H_{0,ate}^{zero}$ is consistent and has correct size asymptotically. Again, we would like to point out that although we adopt the instrument function approach in Andrews and Shi (2013, 2014) other testing procedures developed in the moment equality literature (Bierens, 1982, 1990, Bierens and Ploberger, 1997, and Whang, 2000, 2001, etc.) could also be potentially modified and used. As is discussed earlier, we adopt the instrument function approach because the resulting test statistic requires the same estimation strategy as in classic RD regressions.

3.3 Testing if the Treatment Effect is Heterogenous

To test for treatment effect heterogeneity, we define the hypotheses as

$$\begin{aligned} &H_{0,ate}^{hetero} : CATE(x) = \gamma, \ \forall\ x \in \mathcal{X}_c \text{ and for some } \gamma \in R, \\ &H_{1,ate}^{hetero} : \ H_{0,ate}^{hetero} \text{ does not hold.} \end{aligned} \tag{3.9}$$

If $CATE(x) = \gamma$ for all $x \in \mathcal{X}_c$ and for some $\gamma \in R$, then the equality would hold with $\gamma = ATE = \nu((\mathbf{0}, 1))$. Then $H_{0,ate}^{hetero}$ would imply that $\nu(\ell) = p(\ell) \cdot \nu((\mathbf{0}, 1))$, where $p(\ell) = E[g_\ell(X_i)|Z_i = c]$ is the conditional probability of $X_i \in C_\ell$. So the hypotheses in (3.9) are equivalent to

$$\begin{aligned} H_{0,ate}^{hetero} &: \ \nu_{hetero,ate}(\ell) = \nu(\ell) - \nu((\mathbf{0}, 1)) \cdot p(\ell) = 0, \ \forall\ \ell \in \mathcal{L}, \\ H_{1,ate}^{hetero} &: \ \nu_{hetero,ate}(\ell) = \nu(\ell) - \nu((\mathbf{0}, 1)) \cdot p(\ell) \neq 0, \ \text{for some } \ell \in \mathcal{L}. \end{aligned} \tag{3.10}$$

When $\ell = (\mathbf{0}, 1)$, $\nu_{hetero,ate}(\ell)$ degenerates to zero. For smaller cubes, $\nu_{hetero,ate}(\ell)$ examines whether the ATE among individuals with characteristic values belonging to C_ℓ is equal to the population ATE multiplied by the proportion of such individuals. The following lemma formally summarizes the equivalence result.

Lemma 3.5 *Suppose that Assumption 2.1 holds. Then the hypotheses in (3.9) are equivalent to those in (3.10).*

Let the estimator for $p(\ell)$ be $\hat{p}(\ell)$ such that

$$\hat{p}(\ell) = \frac{\sum_{i=1}^n K(\frac{Z_i - c}{h})[S_{n,2} - S_{n,1}(Z_i - c)]g_\ell(X_i)}{S_{n,0}S_{n,2} - S_{n,1}^2} \equiv \sum_{i=1}^n w_{ni} \cdot g_\ell(X_i),$$

where

$$w_{ni} = \frac{K(\frac{Z_i - c}{h})[S_{n,2} - S_{n,1}(Z_i - c)]}{S_{n,0}S_{n,2} - S_{n,1}^2}, \ \ S_{n,j} = \sum_i K\left(\frac{Z_i - c}{h}\right)(Z_i - c)^j, \ \text{for } j = 0, 1, ...$$

Let $\phi_{p,ni}(\ell) = \sqrt{nh}\Big(w_{ni}(g_\ell(X_i) - p(\ell))\Big)$. Similar to Lemma 3.2, we can show that

$$\left|\sqrt{nh}(\hat{p}(\ell) - p(\ell)) - \sum_{i=1}^n \phi_{p,ni}(\ell)\right| = o_p(1), \tag{3.11}$$

uniformly over $\ell \in \mathcal{L}$. Let $\hat{\nu}_{hetero,ate}(\ell) = \hat{\nu}(\ell) - \hat{\nu}((\mathbf{0}, 1)) \cdot \hat{p}(\ell)$ be the estimator of $\nu_{hetero,ate}(\ell)$. Let $\phi_{ate,ni}^{hetero}(\ell) = \phi_{\nu,ni}(\ell) - p(\ell) \cdot \phi_{\nu,ni}((\mathbf{0}, 1)) - \nu((\mathbf{0}, 1)) \cdot \phi_{p,ni}(\ell)$. It is easy to show that

$$\left|\sqrt{nh}(\hat{\nu}_{hetero,ate}(\ell) - \nu_{hetero,ate}(\ell)) - \sum_{i=1}^n \phi_{ate,ni}^{hetero}(\ell)\right| = o_p(1) \tag{3.12}$$

uniformly over $\ell \in \mathcal{L}$. We give the proof in the online appendix.

Let $\hat{\phi}^{hetero}_{ate,ni}(\ell) = \hat{\phi}_{\nu,ni}(\ell) - \hat{p}(\ell) \cdot \hat{\phi}_{\nu,ni}((\mathbf{0},1)) - \hat{\nu}((\mathbf{0},1)) \cdot \hat{\phi}_{p,ni}(\ell)$ be the estimated influence function with $\hat{\phi}_{p,ni}(\ell) = \sqrt{nh}\Big(w_{ni}(g_\ell(X_i) - \hat{p}(\ell))\Big)$. Define the KS type test statistic as

$$\widehat{S}^{hetero}_{ate} = \sqrt{nh} \sup_{\ell \in \mathcal{L}} \frac{\left|\hat{\nu}_{hetero,ate}(\ell)\right|}{\hat{\sigma}^{hetero}_{ate,\epsilon}(\ell)},$$

where $\hat{\sigma}^{hetero}_{ate,\epsilon}(\ell) = \sqrt{\max\left\{\sum_{i=1}^n \left(\hat{\phi}^{hetero}_{ate,ni}(\ell)\right)^2, \epsilon \cdot \hat{\sigma}^2_{\nu,n}((\mathbf{0},1))\right\}}$ for some small positive ϵ. Again, $\hat{\sigma}^2_{\nu,n}((\mathbf{0},1))$ is used in the definition of $\hat{\sigma}^{hetero}_{ate,\epsilon}(\ell)$ to obtain a scale invariant test statistic. Define the simulated process $\widehat{\Phi}^{hetero,u}_{n,ate}(\ell) = \sum_{i=1}^n U_i \cdot \hat{\phi}^{hetero}_{ate,ni}(\ell)$. For significance level $\alpha < 1/2$, define the simulated critical value $\hat{c}^{hetero}_{n,ate}(\alpha)$ as

$$\hat{c}^{hetero}_{n,ate}(\alpha) = \sup\left\{q \Big| P^u\Big(\sup_{\ell \in \mathcal{L}} \frac{\left|\widehat{\Phi}^{hetero,u}_{n,ate}(\ell)\right|}{\hat{\sigma}^{hetero}_{ate,\epsilon}(\ell)} \le q\Big) \le 1 - \alpha\right\}.$$

Let the decision rule be: "Reject $H^{hetero}_{0,ate}$ if $\widehat{S}^{hetero}_{ate} > \hat{c}^{hetero}_{n,ate}(\alpha)$." The following theorem summarizes the asymptotic properties of the proposed heterogeneity test.

Theorem 3.4 *Under Assumption 2.1 and Assumptions 3.1-3.3, when $\alpha < 1/2$, we have*

(1) under $H^{hetero}_{0,ate}$, $\lim_{n\to\infty} P(\widehat{S}^{hetero}_{ate} > \hat{c}^{hetero}_{n,ate}(\alpha)) = \alpha$, and

(2) under $H^{hetero}_{1,ate}$, $\lim_{n\to\infty} P(\widehat{S}^{hetero}_{ate} > \hat{c}^{hetero}_{n,ate}(\alpha)) = 1$.

Note that this heterogeneity test can also be directly applied to test for first stage heterogeneity in a fuzzy RD model as the selection equation in any fuzzy RD model follows a sharp RD design.

4 Testing in Fuzzy RD Design

In this section, we extend the proposed tests to the fuzzy RD design. Similar to the sharp RD case, we are interested in testing the following three null hypotheses:

$$H^{neg}_{0,late} : CLATE(x) \le 0, \ \forall\ x \in \mathcal{X}_c, \tag{4.1}$$

$$H^{zero}_{0,late} : CLATE(x) = 0, \ \forall\ x \in \mathcal{X}_c, \tag{4.2}$$

$$H^{hetero}_{0,late} : CLATE(x) = \tau, \ \forall\ x \in \mathcal{X}_c \text{ and for some } \tau \in R. \tag{4.3}$$

Recall that $CLATE(x) = \frac{\lim_{z \searrow c} E[Y_i|X_i=x,Z_i=z] - \lim_{z \nearrow c} E[Y_i|X_i=x,Z_i=z]}{E[T_i(1)-T_i(0)|X_i=x,Z_i=c]}$. Since Assumption 2.2.(iv) requires $CLATE$ to have a uniformly positive denominator, the first two hypotheses $H_{0,late}^{neg}$ and $H_{0,late}^{zero}$ will hold if and only if the numerator, $\lim_{z \searrow c} E[Y_i|X_i = x, Z_i = z] - \lim_{z \nearrow c} E[Y_i|X_i = x, Z_i = z]$, is uniformly negative or uniformly zero, respectively. In other words, these two hypotheses can be tested by applying the testing procedures developed in Section 3.

For the third hypothesis, the null $CLATE(x) = \tau$ holds for all $x \in \mathcal{X}_c$ and for some $\tau \in R$ if and only if $CLATE(x) = LATE$ for all $x \in \mathcal{X}_c$. Both the $CLATE(x)$ and the $LATE$ could be estimated using local linear regression methods. However, developing heterogeneity tests relying on plug-in estimators of the $LATE$ or the $CLATE$ are not ideal. As is discussed in Feir et al. (2015), when the sample size or the proportion of compliers is small the $LATE$ estimator can have a Cauchy-type finite sample distribution due to the random denominator problem, analogous to the concerns raised in the weak IV literature (see, e.g., Staiger and Stock, 1997). This problem is even worse with the $CLATE$ as the effect conditions on both the running variable Z_i and the additional covariate X_i. The RD model could also have a heterogeneous first stage which can lead to low first stage compliance rate for some subpopulations. In the rest of the session, we look at null transformations that can avoid direct use of the $LATE$ or $CLATE$.

Let

$$\mu(\ell) = \lim_{z \searrow c} E[g_\ell(X_i)T_i|Z_i = z] - \lim_{z \nearrow c} E[g_\ell(X_i)T_i|Z_i = z].$$

It is clear that the $LATE = \nu((\mathbf{0},1))/\mu((\mathbf{0},1))$ and that $\nu(\ell)/\mu(\ell)$ is the local average treatment effect for individuals with $X_i \in C_\ell$. In the online appendix, we show that the null hypothesis in (4.3) is equivalent to

$$H_{0,late}^{hetero}: \ \nu_{hetero,late}(\ell) = \nu(\ell) \cdot \mu((\mathbf{0},1)) - \nu((\mathbf{0},1)) \cdot \mu(\ell) = 0, \ \forall \ \ell \in \mathcal{L}. \tag{4.4}$$

Let $\hat{\mu}(\ell)$ be the estimator for $\mu(\ell)$ that is defined in the same way as $\hat{\nu}(\ell)$ except that Y_i is replaced by T_i. Let $\hat{\nu}_{hetero,late}(\ell) = \hat{\nu}(\ell) \cdot \hat{\mu}((\mathbf{0},1)) - \hat{\nu}((\mathbf{0},1)) \cdot \hat{\mu}(\ell)$ be the estimator for $\nu_{hetero,late}(\ell)$. Let $\phi_{\mu,ni}(\ell)$ be the influence function for $\sqrt{nh}(\hat{\mu}(\ell) - \mu(\ell))$ that is defined in the same way as $\phi_{\nu,ni}(\ell)$ except that Y_i is replaced by T_i, and let $\hat{\phi}_{\mu,ni}(\ell)$ be its estimator. Let $\phi_{late,ni}^{hetero}(\ell) = \mu((\mathbf{0},1)) \cdot \phi_{\nu,ni}(\ell) + \nu(\ell) \cdot \phi_{\mu,ni}((\mathbf{0},1)) - \nu((\mathbf{0},1)) \cdot \phi_{\mu,ni}(\ell) - \mu(\ell) \cdot \phi_{\nu,ni}((\mathbf{0},1))$

and $\hat{\phi}^{hetero}_{late,ni}(\ell) = \hat{\mu}((\mathbf{0},1))\cdot\hat{\phi}_{\nu,ni}(\ell)+\hat{\nu}(\ell)\cdot\hat{\phi}_{\mu,ni}((\mathbf{0},1))-\hat{\nu}((\mathbf{0},1))\cdot\hat{\phi}_{\mu,ni}(\ell)-\hat{\mu}(\ell)\cdot\hat{\phi}_{\nu,ni}((\mathbf{0},1))$. Let $\left(\hat{\sigma}^{hetero}_{late,n}(\ell)\right)^2 = \sum_{i=1}^n \left(\hat{\phi}^{hetero}_{late,ni}(\ell)\right)^2$. Define the test statistic for $H^{hetero}_{0,late}$ as

$$\widehat{S}^{hetero}_{late} = \sqrt{nh}\sup_{\ell\in\mathcal{L}} \frac{\left|\hat{\nu}_{hetero,late}(\ell)\right|}{\hat{\sigma}^{hetero}_{late,\epsilon}(\ell)},$$

where $\hat{\sigma}^{hetero}_{late,\epsilon}(\ell) = \sqrt{\max\left\{\left(\hat{\sigma}^{hetero}_{late,n}(\ell)\right)^2, \epsilon\cdot\hat{\sigma}^2_{\nu,n}((\mathbf{0},1))\right\}}$ for some small positive ϵ. Define the simulated process $\widehat{\Phi}^{hetero,u}_{n,late}(\ell)$ as $\widehat{\Phi}^{hetero,u}_{n,late}(\ell) = \sum_{i=1}^n U_i\cdot\hat{\phi}^{hetero}_{late,ni}(\ell)$.

For significance level $\alpha < 1/2$, define the simulated critical value $\hat{c}^{hetero}_{n,late}(\alpha)$ as

$$\hat{c}^{hetero}_{n,late}(\alpha) = \sup\left\{q\Big|P^u\Big(\sup_{\ell\in\mathcal{L}}\frac{\left|\widehat{\Phi}^{hetero,u}_{n,late}(\ell)\right|}{\hat{\sigma}^{hetero}_{late,\epsilon}(\ell)}\le q\Big)\le 1-\alpha\right\}.$$

Finally, the decision rule would be: "Reject $H^{hetero}_{0,late}$ if $\widehat{S}^{hetero}_{late} > \hat{c}^{hetero}_{n,late}(\alpha)$." Again, the proposed test for $H^{hetero}_{0,late}$ controls size asymptotically and is consistent. We omit the details of the size and power properties in the interest of space.

5 Simulations

In this section, we carry out Monte Carlo simulations. First, we investigate the small sample size and power performance of the proposed tests using data generating processes (DGPs) estimated from the dataset in the empirical section. Second, we design some special DGPs to demonstrate 1) the size and power performances of the proposed uniform sign test based on the LFC and GMS critical values, 2) the size distortion of the interaction term method and the subsample regression method popular in the applied RD literature for heterogeneity analysis, and 3) the small sample performance of the proposed tests when the DGP introduces larger finite sample bias in local linear estimation.

For all DGPs, the running variable Z, the additional control X, and the error term u are generated following

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

The outcome Y and the treatment decision T are DGP specific. With each DGP, 5,000 simulation samples are drawn unless otherwise noted. In each test, the bootstrap critical value is calculated using 1,000 bootstrap simulations.

In the simulation and empirical sections of the paper, we follow the RD literature and use the triangular kernel (i.e., $K(u) = (1 - |u|) \cdot 1(|u| \leq 1)$) for all boundary local linear estimators. We also set the bandwidth to $h_{CCT} \times n^{1/5-1/k}$, where h_{CCT} is the robust bandwidth following Calonico et al. (2014) (CCT), and the multiplicative factor $n^{1/5-1/k}(k < 5)$ is used to obtain the under-smoothed bandwidth required in our testing procedure. To make sure that the simulation results are not sensitive to the under-smoothing factor, we report in all simulation tables results with three different k choices: 4.25, 4.5 and 4.75. The cubes defined in equation (3.2) have side-lengths $1/q$ for $q = 1, ..., Q$. We use as a benchmark $Q = 10$ which includes a total of 55 overlapping intervals. Andrews and Shi (2013) suggest choosing Q such that the smallest cubes have expected sample size around 10 to 20. With this benchmark Q choice, when $n = 1,000$ the smallest cubes of each local linear regression in DGPs 1-4 have expected effective sample sizes ranging from 16 to 21.[6] In Tables 1-3 we also report robustness checks with $Q = 7$ and 13.

Models of Y and T in the DGPs are estimated with the dataset in the empirical section. DGP 1 models a sharp RD design with a homogeneous zero effect of the treatment variable T. To model the outcome Y, we first normalize the domain of the additional control of interest (i.e., peer transition score) to $[0, 1]$ and then regress the outcome (i.e., Baccalaureate exam score) on the running variable (i.e., transition score), the additional control of interest, as well as their interaction term and second order polynomial terms. DGP 2 models a sharp RD design with a heterogeneous treatment effect. The outcome equation is obtained by fitting the same regression model described above separately for the subsamples to the left and the right of the cutoff value (i.e., zero) of the running variable. DGPs 3 and 4 are fuzzy RD models with the same outcome equations as in DGPs 1 and 2, respectively. The first stage equation is generated such that the treatment dummy is zero for all data to the left of the cutoff, and follows a Probit model estimated with the empirical dataset for all data to the right of the cutoff.

Models of Y and T for DGPs 1-4 are specified in below. The models for the outcome variable Y are also visualized in Figure 1.

[6] The calculation of effective sample size takes into account the distribution of Z as well as the average bandwidths of the local linear regressions among the 5,000 simulations.

DGP 1: Sharp RD, Homogeneous Zero Effect

$$Y = -0.555 - 0.553X + 0.581Z + 0.060XZ - 0.058Z^2 + 1.074X^2 + 0.1u;$$
$$T = 1(Z > 0).$$

DGP 2: Sharp RD, Heterogeneous Effects

$$Y = \begin{cases} -0.755 - 0.254X + 0.742Z - 0.219XZ - 0.063Z^2 + 1.175X^2 + 0.1u & \text{if } Z \geq 0, \\ -0.607 - 0.220X + 0.386Z + 0.288XZ + 0.204Z^2 + 0.469X^2 + 0.1u & \text{if } Z < 0; \end{cases}$$
$$T = 1(Z > 0).$$

DGP 3: Fuzzy RD, Homogeneous Zero Effect

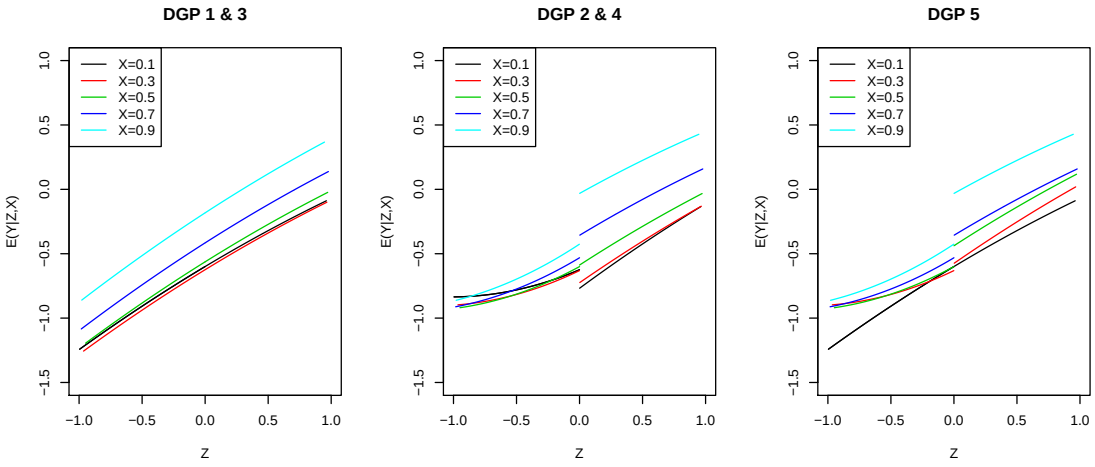
$$Y \sim \ DGP\ 1;$$
$$T = \begin{cases} 1(0.596 - 2.103X + 0.128Z + 0.352XZ + 0.013Z^2 + 2.454X^2 + u > 0) & \text{if } Z \geq 0, \\ 0 & \text{if } Z < 0. \end{cases}$$

DGP 4: Fuzzy RD, Heterogeneous Effects

$$Y \sim \ DGP\ 2;$$
$$T = \begin{cases} 1(0.596 - 2.103X + 0.128Z + 0.352XZ + 0.013Z^2 + 2.454X^2 + u > 0) & \text{if } Z \geq 0, \\ 0 & \text{if } Z < 0. \end{cases}$$

Table 1 reports the simulation results of DGPs 1 and 2, including the uniform sign test ($H_0 : CATE(x) \leq 0,\ \forall x \in [0,1]$) and the heterogeneity test ($H_0 : CATE(x) = ATE,\ \forall x \in [0,1]$). Simulation results of the overall significance test ($H_0 : CATE(x) = 0,\ \forall x \in [0,1]$) are omitted as they are highly similar to the results of the uniform sign test. Table 1 also reports the results of the standard mean test ($H_0 : ATE = 0$) for comparison. Simulation results of DGP 1 show that the proposed tests control size well. Even when the sample size is small with $n = 1,000$, the rejection rate is controlled under 5.5%, which is close to the 5% significance level. The simulation results of DGP 2 show that the proposed tests have good power performance with the rejection rate going to one as the sample size increases. The reported rejection rates are somewhat dependent on the bandwidth choice, which is common to all kernel based tests. In the empirical application we also report testing results with the same three under-smoothing factors. We find that our empirical findings are very robust to the bandwidth choice. Finally,

Figure 1: The Data Generating Processes: DGPs 1-5



Note: The left graph plots the outcome equation of DGP 1 & 3. The middle graph plots the outcome equation of DGP 2 & 4. The right graph plots the outcome equation of DGP 5. All figures plot $E(Y|Z,X)$ against the whole domain of Z and five difference values of X.

simulation results in Table 1 show that the size and power performance of the proposed tests are robust to the choice of Q.

Table 2 reports the simulation results of the heterogeneity test ($H_0 : CLATE(x) = LATE,\ \forall x \in [0,1]$) for DGPs 3 and 4. Simulation results of the standard mean test and the uniform sign test are omitted because they are exactly the same as the results for DGPs 1 and 2 reported in Table 1. This is because testing the sign and significance of $LATE$ or $CLATE$ under fuzzy RD designs is the same as testing the sign or significance of its numerator (see the identification result in equation (2.2)), which is the same in DGPs 3 and 4 as in DGPs 1 and 2, respectively.

In the above simulation experiments, we have been using the LFC critical value in the uniform sign test. Next, we illustrate the size and power performance of the uniform sign test with the GMS critical value. The right graph in Figure 1 visualizes the outcome equation in DGP 5. We test both the null hypothesis of a uniformly non-positive effect ($H_0 : CATE(x) \leq 0,\ \forall x \in [0,1]$) and the null of a uniformly non-negative effect ($H_0 : CATE(x) \geq 0,\ \forall x \in [0,1]$).

Table 1: Proposed Tests Under Sharp RD

	$H_0 : ATE = 0$			$H_0 : CATE(\cdot) \leq 0$			$H_0 : CATE(\cdot) = ATE$		
	k=4.25	k=4.5	k=4.75	k=4.25	k=4.5	k=4.75	k=4.25	k=4.5	k=4.75
Panel A: $Q = 10$									
DGP 1: Homogeneous Zero Effect									
n=1000	0.053	0.056	0.059	0.054	0.054	0.053	0.053	0.052	0.048
n=2000	0.057	0.059	0.060	0.055	0.053	0.054	0.051	0.054	0.054
n=4000	0.050	0.050	0.053	0.047	0.049	0.049	0.047	0.048	0.046
n=8000	0.052	0.056	0.057	0.048	0.052	0.051	0.049	0.048	0.051
DGP 2: Heterogeneous Treatment Effect									
n=1000	0.279	0.301	0.324	0.279	0.318	0.357	0.178	0.191	0.205
n=2000	0.497	0.527	0.561	0.616	0.672	0.725	0.323	0.362	0.395
n=4000	0.745	0.779	0.809	0.935	0.958	0.972	0.630	0.686	0.736
n=8000	0.924	0.946	0.958	0.999	1.000	1.000	0.931	0.956	0.974
Panel B: $Q = 7$									
DGP 1: Homogeneous Zero Effect									
n=1000	0.053	0.056	0.059	0.054	0.054	0.054	0.055	0.055	0.050
n=2000	0.057	0.059	0.060	0.055	0.053	0.054	0.051	0.055	0.053
n=4000	0.050	0.050	0.053	0.048	0.048	0.049	0.048	0.048	0.046
n=8000	0.052	0.056	0.057	0.048	0.052	0.051	0.050	0.049	0.052
DGP 2: Heterogeneous Treatment Effect									
n=1000	0.279	0.301	0.324	0.286	0.325	0.365	0.184	0.196	0.212
n=2000	0.497	0.527	0.561	0.623	0.679	0.731	0.329	0.369	0.403
n=4000	0.745	0.779	0.809	0.937	0.960	0.974	0.634	0.694	0.742
n=8000	0.924	0.946	0.958	0.999	1.000	1.000	0.935	0.957	0.976
Panel C: $Q = 13$									
DGP 1: Homogeneous Zero Effect									
n=1000	0.053	0.056	0.059	0.054	0.054	0.053	0.052	0.051	0.047
n=2000	0.057	0.059	0.060	0.054	0.053	0.053	0.051	0.053	0.054
n=4000	0.050	0.050	0.053	0.047	0.048	0.048	0.048	0.048	0.046
n=8000	0.052	0.056	0.057	0.048	0.052	0.051	0.049	0.048	0.051
DGP 2: Heterogeneous Treatment Effect									
n=1000	0.279	0.301	0.324	0.277	0.316	0.354	0.177	0.190	0.204
n=2000	0.497	0.527	0.561	0.613	0.670	0.723	0.321	0.360	0.393
n=4000	0.745	0.779	0.809	0.934	0.958	0.972	0.627	0.685	0.735
n=8000	0.924	0.946	0.958	0.999	1.000	1.000	0.928	0.956	0.974

Note: Reported are rejection proportions among 5,000 simulations where all tests are carried out using the 5% significance level. The uniform sign test ($H_0 : CATE(\cdot) \leq 0$) is carried out using the LFC critical value. For each test the simulated critical value is calculated with 1,000 bootstrap repetitions.

Table 2: Heterogeneity Test under Fuzzy RD

	$Q=7$			$Q=10$			$Q=13$		
	k=4.25	k=4.5	k=4.75	k=4.25	k=4.5	k=4.75	k=4.25	k=4.5	k=4.75
DGP 3: Homogeneous Zero Effect									
n=1000	0.047	0.047	0.046	0.045	0.045	0.044	0.045	0.045	0.043
n=2000	0.049	0.052	0.050	0.050	0.052	0.052	0.049	0.052	0.051
n=4000	0.047	0.047	0.045	0.046	0.047	0.045	0.046	0.047	0.044
n=8000	0.049	0.049	0.051	0.049	0.047	0.052	0.049	0.048	0.052
DGP 4: Heterogeneous Treatment Effect									
n=1000	0.151	0.169	0.182	0.145	0.167	0.177	0.146	0.165	0.175
n=2000	0.278	0.310	0.337	0.272	0.303	0.330	0.271	0.301	0.329
n=4000	0.534	0.589	0.644	0.529	0.583	0.635	0.527	0.581	0.634
n=8000	0.867	0.906	0.932	0.864	0.902	0.930	0.862	0.901	0.930

Note: Reported are rejection proportions of the Fuzzy RD uniform sign test ($H_0 : CATE(x) \leq 0,\ \forall x \in [0,1]$) among 5,000 simulations. All tests are carried out using the 5% significance level and the LFC critical value. For each test, the simulated critical value is calculated with 1,000 bootstrap repetitions.

DGP 5: Sharp RD, Mixture of Zero and Positive Effects

$$\text{When } X < 0.3, Y \sim \ DGP\ 1;\ \text{when } X \geq 0.6, Y \sim \ DGP\ 2;\ \text{when } 0.3 \leq X < 0.6;$$

$$Y = \begin{cases} -0.755 - 0.254X + 0.742Z - 0.219XZ - 0.063Z^2 + 1.175X^2 + 0.1u & \text{if } Z \geq 0, \\ -0.607 - 0.220X + 0.386Z + 0.288XZ + 0.204Z^2 + 0.469X^2 + 0.1u & \text{if } Z < 0. \end{cases}$$

Under DGPs 1 and 3, the uniform sign test using the LFC critical value controls size well and has rejection rates very close to the nominal rate of 5%. The reason is that the true treatment effects are uniformly zero for all values of X, which means that the LFC always holds for the tests. Under DGP 5, the null hypothesis of a non-negative effect is true, but the LFC holds only when $X < 0.3$. Since the LFC does not hold at all times, using the LFC critical value will be conservative while using the GMS critical value should bring the rejection rate closer to 5%. Meanwhile, the null hypothesis of a non-positive effect is false under DGP 5, and the GMS critical value should improve the power performance of the test. This is exactly what is shown in Table 3.

DGPs 6 and 7 are designed to facilitate the comparison between the proposed heterogeneity test (Hetero) and two popular tests in the applied RD literature: the interaction term method (Hetero-INT) and the subsample regression method (Hetero-SUB). The Hetero-INT test is carried out by testing the slope coefficient on the interaction term

Table 3: Uniform Sign Tests Using the LFC V.s. the GMS Critical Values

	LFC						GMS					
	$H_0 : CATE(\cdot) \geq 0$			$H_0 : CATE(\cdot) \leq 0$			$H_0 : CATE(\cdot) \geq 0$			$H_0 : CATE(\cdot) \leq 0$		
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: $Q = 10$												
n=1000	0.040	0.041	0.040	0.800	0.841	0.880	0.048	0.049	0.049	0.813	0.854	0.888
n=2000	0.040	0.040	0.039	0.982	0.988	0.992	0.047	0.048	0.045	0.983	0.989	0.993
n=4000	0.033	0.033	0.033	1.000	1.000	1.000	0.039	0.040	0.041	1.000	1.000	1.000
n=8000	0.031	0.030	0.029	1.000	1.000	1.000	0.038	0.038	0.035	1.000	1.000	1.000
Panel B: $Q = 7$												
n=1000	0.029	0.031	0.032	0.659	0.716	0.761	0.038	0.038	0.040	0.672	0.729	0.769
n=2000	0.032	0.030	0.030	0.933	0.957	0.971	0.039	0.039	0.038	0.934	0.959	0.972
n=4000	0.023	0.024	0.022	0.998	0.999	1.000	0.031	0.030	0.030	0.999	0.999	1.000
n=8000	0.023	0.022	0.023	1.000	1.000	1.000	0.029	0.029	0.031	1.000	1.000	1.000
Panel C: $Q = 13$												
n=1000	0.040	0.041	0.040	0.798	0.839	0.877	0.047	0.048	0.048	0.811	0.852	0.886
n=2000	0.039	0.038	0.039	0.981	0.988	0.992	0.047	0.047	0.045	0.982	0.989	0.993
n=4000	0.033	0.033	0.033	1.000	1.000	1.000	0.039	0.040	0.040	1.000	1.000	1.000
n=8000	0.033	0.033	0.032	1.000	1.000	1.000	0.038	0.038	0.039	1.000	1.000	1.000

Note: Reported are rejection proportions among 5,000 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.

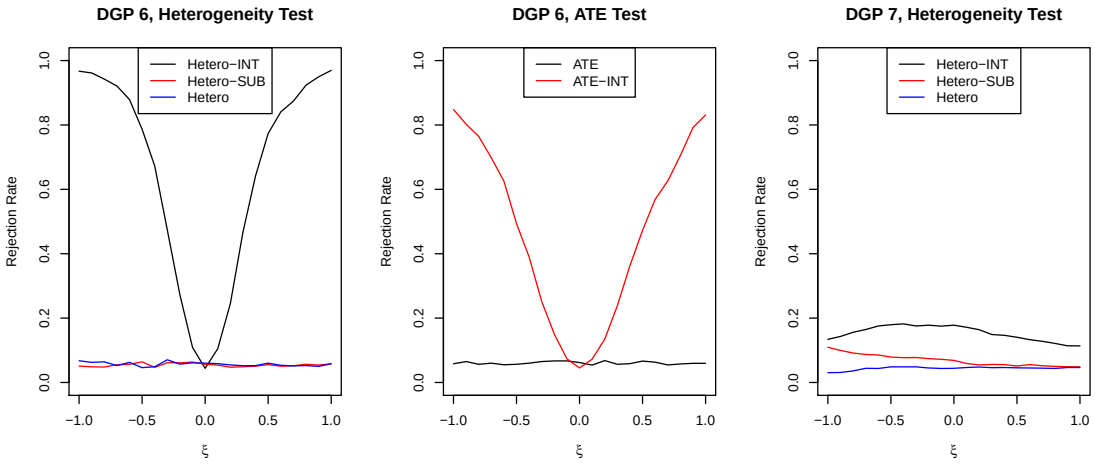
$X1(Z > 0)$ in the linear regression of Y on X, Z, $1(Z > 0)$, $X1(Z > 0)$, and $Z1(Z > 0)$, using data inside the estimation window determined by the CCT bandwidth. The Hetero-SUB test is carried out by testing whether subsample average treatment effects of any of the five subsamples with $X = [0, 0.2], X = (0.2, 0.4], X = (0.4, 0.6], X = (0.6, 0.8], X = (0.8, 1]$ are different from the true population ATE. The Hetero-SUB test adjusts for multiple testing using the Bonferroni method. The interaction term method is expected to over-reject when the model is not linear. The subsample regression method is expected to have a sizable over-rejection rate when the sample size is small and/or the first-stage take-up rate is low.

DGP 6: Sharp RD, Homogeneous Zero Effect

$$\begin{aligned} Y =& -0.483 - 1.376X + 0.301Z \\ &+ \xi(0.112XZ + 0.194Z^2 + 3.234X^2 - 0.295XZ^2 + 0.469XZ^2 - 1.548X^3 - 0.021Z^3) \\ &+ 0.1u; \\ T =& 1(Z > 0). \end{aligned}$$

Under DGP 6, the treatment effect is homogenous and zero, and the control parameter ξ determines the degree of model misspecification. The left graph of Figure 2 summarizes the size property of the three tests. We notice that the parametric Hetero-INT test controls size at 5% only when $\xi = 0$ and the linear regression model is correctly specified. In contrast, the Hetero and Hetero-SUB tests control size well irrespective of ξ because both tests are nonparametric. In addition to the heterogeneity tests, we also report in the middle graph of Figure 2 rejection rates of the ATE significance test based on ATE estimators from the classic RD regression (ATE) and the interaction term method (ATE-INT). We see from the graph that the ATE test based on the interaction term method also over-rejects severely unless $\xi = 0$, which further supports our recommendation against the use of the interaction term method in applied work.

Figure 2: Performance of Naive and Proposed Testing Methods



Note: Reported are rejection proportions among 1,000 simulations. All tests are carried out using the 5% significance level and the simulated critical values calculated with 1,000 bootstrap repetitions. The sample size is $1,000$. The Hetero test is the proposed heterogeneity test with $k = 4.5$ and $Q = 10$. The Hetero-INT test is the heterogeneity test using the interaction term method. The Hetero-SUB test is the heterogeneity based on subsample RD regressions. Details of the Hetero-INT and the Hetero-SUB tests are described in the main text.

DGP 7: Fuzzy RD, Homogeneous Positive Effect

$$Y = \begin{cases} -0.755 + 0.742Z - 0.063Z^2 + 0.1u & \text{if } Z \geq 0, \\ -0.607 + 0.386Z + 0.204Z^2 + 0.1u & \text{if } Z < 0; \end{cases}$$

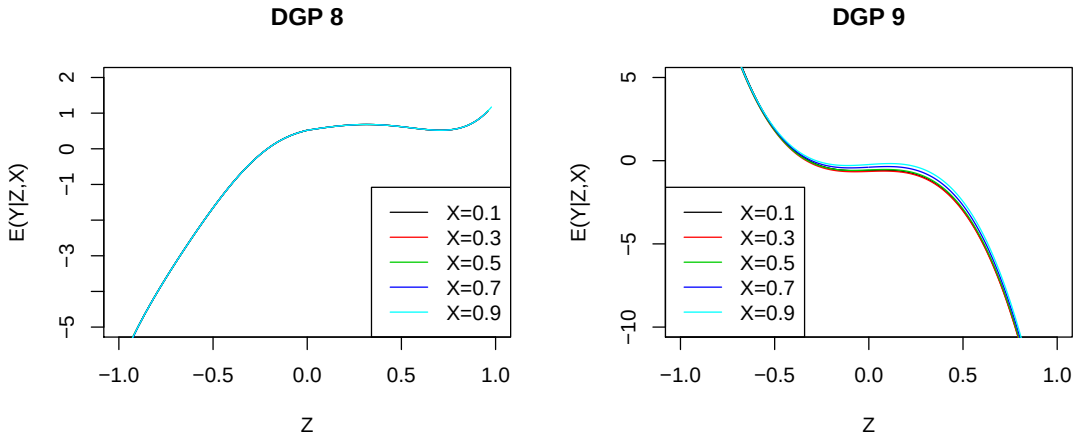
$$T = \begin{cases} 1(\xi \cdot 0.596 - 2.103X + 0.128Z + 0.352XZ + 0.013Z^2 + 2.454X^2 + u > 0) & \text{if } Z \geq 0, \\ 0 & \text{if } Z < 0. \end{cases}$$

DGP 7 is modified from DGP 2 by suppressing the role of the additional covariate X in the outcome equation so that the model has a homogeneous and positive effect. The control parameter ξ in the selection equation determines the first stage take-up rate. The smaller the value of ξ, the weaker the first stage.[7] As is shown in the right graph of Figure 2, the proposed heterogeneity test has excellent size control while the Hetero-SUB test has sizable over-rejection when the first-stage is weak. The parametric Hetero-INT test again over-rejects because of model misspecification.

Last but not least, we examine the small sample performance of the proposed tests

[7]The first-stage take-up rate for different X values ranges from 0.15-0.3 when $\xi = -1$ to 0.55-0.75 when $\xi = 1$.

Figure 3: The Data Generating Processes: DGPs 8-9



Note: DGP 8 is taken from Calonico et al. (2014). DGP 9 is modified from a data-driven model estimated from the dataset of the empirical section.

when the DGP has an asymmetric and exaggerated curvature pattern around the cut-off value of the running variable. DGP 8 is taken from Calonico et al. (2014), which is specifically designed to show the importance of bias correction. In this paper we do not use the bias correction technique but instead require an under-smoothed bandwidth to avoid having nuisance bias terms in the limiting distributions of local linear estimators. Thus this DGP from Calonico et al. (2014) should serve as a good example to examine the small sample performance of our testing procedure when the DGP does not favor the under-smoothing technique that we employ.

DGP 8: Sharp RD, Homogeneous Zero Effect, Exaggerated Curvature

$$Y = \begin{cases} 0.52 + 0.84Z - 0.3Z^2 - 2.4Z^3 - 0.9Z^4 + 3.56Z^5 + 0.1u & \text{if } Z \geq 0, \\ 0.52 + 1.27Z - 3.59Z^2 + 14.15Z^3 + 23.69Z^4 + 11.36Z^5 + 0.1u & \text{if } Z < 0; \end{cases}$$

$T = 1(Z > 0)$.

The left graph in Figure 3 illustrates the model in DGP 8 and Table 4 reports the simulation results. For results reported in the first nine columns, we only see slight over-rejection with rejection rates always under 7%. The rejection rate also gets quite close to the 5% nominal rate when the sample size gets larger. However, when the GMS critical

Table 4: Uniform Sign Test Using the LFC and the GMS Critical Values

	$H_0 : CATE(\cdot) \leq 0$ (LFC)			$H_0 : CATE(\cdot) = LATE$			$H_0 : CATE(\cdot) \leq 0$ (GMS)		
	k=4.25	k=4.5	k=4.75	k=4.25	k=4.5	k=4.75	k=4.25	k=4.5	k=4.75
DGP 8: Homogeneous Zero Effect, Exaggerated Curvature									
n=1000	0.062	0.063	0.066	0.059	0.058	0.057	0.081	0.085	0.086
n=2000	0.057	0.056	0.055	0.052	0.055	0.052	0.074	0.073	0.071
n=4000	0.057	0.059	0.060	0.054	0.057	0.056	0.077	0.075	0.075
n=8000	0.058	0.061	0.059	0.055	0.057	0.056	0.076	0.073	0.072
DGP 9: Homogeneous Zero Effect, Exaggerated Curvature									
n=1000	0.062	0.062	0.063	0.058	0.057	0.055	0.075	0.074	0.073
n=2000	0.061	0.057	0.057	0.052	0.052	0.050	0.073	0.069	0.070
n=4000	0.051	0.052	0.053	0.049	0.049	0.047	0.063	0.061	0.062
n=8000	0.053	0.053	0.053	0.051	0.053	0.056	0.062	0.061	0.062

Note: Reported are rejection proportions among 5,000 simulations. All tests are carried out using the 5% significance level and the simulated critical values calculated with 1,000 bootstrap repetitions. All tests in this table use $Q = 10$.

value is used for the uniform sign test the rejection rate gets higher – close to 9% when $n = 1,000$ – although it then steadily decreases as the sample size gets larger.

DGP 9: Sharp RD, Homogeneously Zero Effect, Exaggerated Curvature

$$Y = \begin{cases} -0.905 + 0.742X - 0.254Z - 0.219XZ - 0.063Z^2 + 1.175X^2 + 0.1u & \text{if } Z \geq 0, \\ -0.607 + 0.386X - 0.220Z + 0.288XZ + 0.204Z^2 + 0.469X^2 + 0.1u & \text{if } Z < 0; \end{cases}$$

$T = 1(Z > 0)$.

To further confirm the findings of DGP 8, we design a new DGP that modifies a higher-order polynomial model estimated with the empirical dataset. DGP 9 also has exaggerated curvature that is asymmetric around the cut-off of the running variable, as is shown in the right graph of Figure 3. Similar to DGP 8, we only observe very mild over-rejection except for the uniform sign test with the GMS critical value. The over-rejection problem again improves with the sample size.

In summary, we conclude that our proposed tests have very good small sample performance. When the underlying RD model has excessive curvature and the sample size is small, using the GMS critical value for the uniform sign test might lead to moderate over-rejection. In such cases, the LFC critical value is recommended.

6 The Heterogeneous Effect of Going to a Better High School

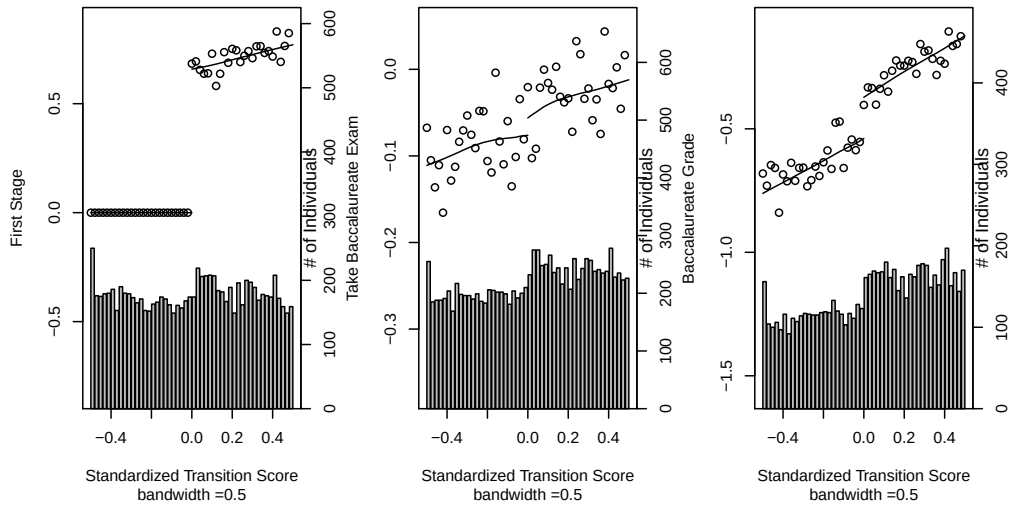
In Romania, a typical elementary school student takes a nationwide test in the last year of school (8th grade) and applies to a list of high schools and tracks. The admission decision is entirely dependent on the student's transition score, an average of the student's performance on the nationwide test and grade point average, and preference for schools. A student with a transition score above a school's cutoff is admitted to the most selective school for which he or she qualifies. Pop-Eleches and Urquiola (2013) use an administrative dataset from Romania to study the impact of attending a more selective high school. They find that attending a better school significantly improves a student's performance on the Baccalaureate exam, but does not affect the exam take-up rate. They also find that a marginal student attending a more selective high school is more likely to face negative peer interactions and perceive himself as weak.

In this section, we investigate the treatment effect heterogeneity among schools with different peer quality, where peer quality is defined as average admission score of the most selective school in town. In contrast to Pop-Eleches and Urquiola (2013), who find qualitatively similar results across schools in three terciles of the admission score cut-off, we find a clear signal that attending a better high school has a heterogeneous effect on whether a marginal student takes the Baccalaureate exam.

Figure 4 summarizes the classic mean RD results. Note that we restrict our attention to two-school towns because score cutoffs within a town are often quite close and estimation bias might be introduced as a result.[8] In all three graphs the x-axis represents the running variable, which is a student's standardized transition score subtracting the school admission cut-off. The y-axis in the left graph represents the first stage take-up rate which is equal to the proportion of eligible students attending a more selective school. The y-axis in the middle and right graphs represent two different outcomes, the demeaned probability of a student taking the Baccalaureate exam and the demeaned Bac-

[8] In fact, it is easy to prove that if the potential outcome monotonically increases with the running variable and also jumps positively at all discontinuity points, having extra discontinuity points within the estimation window can severely downward bias the ATE estimator.

Figure 4: Pooled Regression Discontinuity Analysis



Notes: Data are from Pop-Eleches and Urquiola (2013). Nonparametric local linear estimations are conducted using a triangular kernel. The bar chart reports the histogram of the standardized running variable, while the circles and lines report the average outcome within each bin and the local linear estimates. The bandwidth is set to 0.5 for all graphs for the purpose of data illustration and cross-comparison.

calaureate exam grade among exam-takers, respectively. Both outcomes are demeaned by subtracting the school fixed effects following Pop-Eleches and Urquiola (2013). Both the middle and the right graphs see a jump in the average outcome at the discontinuity point, although the jump in the exam-taking rate is quite noisy.

Table 5 reports the testing results for the heterogeneity analysis. All tests use the triangular kernel, the undersmoothed CCT bandwidth defined in the simulation section, and the cubes defined in (3.2). In Table 5 the tests are conducted with $Q = 100$.[9]

The testing results are very interesting. As is shown in Figure 4 and Table 5 (first row, columns 1-3), the average effect of attending a better school on the probability of a marginal student taking the Baccalaureate exam is noisy and statistically insignificant. But the additional testing results in Table 5 reveal that 1) we can reject the null of a non-positive effect at about the 1% significance level (first row, columns 4-6), 2) we can reject the null that the effect is non-negative at the 10% significance level when the GMS

[9]Given the large sample size of this empirical application, adopting the recommendation in Andrews and Shi (2013) to select Q is computationally costly. Instead, we report results with $Q = 75$, 100, and 125. We find that our empirical findings are insensitive to the choice of Q.

Table 5: Uniform Sign and Heterogeneity Tests

$H_0: LATE = 0$			$CLATE(\cdot) \leq 0$			$CLATE(\cdot) \geq 0$			$CLATE(\cdot) = LATE$		
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
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Exam-taking Rate (LFC for Uniform Sign Tests)											
0.482	0.350	0.285	0.012	0.007	0.005	0.104	0.116	0.111	0.004	0.001	0.001
Exam Grade (LFC for Uniform Sign Tests)											
0.002	0.003	0.002	0.002	0.002	0.001	0.652	0.687	0.749	0.074	0.103	0.135
Exam-taking Rate (GMS for Uniform Sign Tests)											
-	-	-	0.012	0.007	0.005	0.092	0.099	0.097	-	-	-
Exam Grade (GMS for Uniform Sign Tests)											
-	-	-	0.002	0.002	0.001	0.545	0.579	0.662	-	-	-
First-stage Take-up Rate (LFC for Uniform Sign Tests)											
0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.003	0.003	0.006
First-stage Take-up Rate (GMS for Uniform Sign Tests)											
-	-	-	0.000	0.000	0.000	1.000	1.000	1.000	-	-	-

Notes: Data are from Pop-Eleches and Urquiola (2013). The numbers reported in the table are p-values of various tests. All simulated critical values are calculated with 1,000 bootstrap repetitions. All tests in this table use $Q = 100$.

critical value is used (third row, columns 7-9), and 3) we can reject the null that the effect does not vary with peer quality in the more selective school at the 1% significance level (first row, columns 10-12). Adding up the three pieces of information we conclude that the insignificant LATE on the exam-taking rate results from the cancellation of negative and positive effects among different schools.[10]

In addition, testing results in Table 5 confirm the positive effect of attending a better school on the Baccalaureate exam grade (second and fourth rows, columns 4-9). The results also reveal strong evidence of first stage heterogeneity (fifth row, columns 10-12), which is intuitive as selective schools with higher peer quality are expected to have a higher attendance rate among qualified students.

Table 6 conducts robustness checks with two alternative Q values. The results suggest that the above discussed empirical results are not sensitive to the choice of Q.

[10]Note that the conclusion also holds after adjusting the p-values reported in Table 4 with Holm's step-down method due to extremely small p-values in two out of the three tests carried out.

Table 6: Uniform Sign and Heterogeneity Tests - Alternative Q Values

$H_0 : LATE = 0$			$CLATE(\cdot) \leq 0$			$CLATE(\cdot) \geq 0$			$CLATE(\cdot) = LATE$		
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=75											
Exam-taking Rate (LFC for Uniform Sign Tests)											
0.482	0.350	0.285	0.012	0.007	0.005	0.104	0.116	0.111	0.004	0.001	0.001
Exam Grade (LFC for Uniform Sign Tests)											
0.002	0.003	0.002	0.002	0.002	0.001	0.651	0.686	0.747	0.074	0.103	0.135
First-stage Take-up Rate (LFC for Uniform Sign Tests)											
0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.003	0.003	0.006
Exam-taking Rate (GMS for Uniform Sign Tests)											
-	-	-	0.012	0.007	0.005	0.092	0.099	0.097	-	-	-
Exam Grade (GMS for Uniform Sign Tests)											
-	-	-	0.002	0.002	0.001	0.544	0.578	0.659	-	-	-
First-stage Take-up Rate (GMS for Uniform Sign Tests)											
-	-	-	0.000	0.000	0.000	1.000	1.000	1.000	-	-	-
Q=125											
Exam-taking Rate (LFC for Uniform Sign Tests)											
0.482	0.350	0.285	0.012	0.007	0.005	0.104	0.116	0.111	0.004	0.001	0.001
Exam Grade (LFC for Uniform Sign Tests)											
0.002	0.003	0.002	0.002	0.002	0.001	0.652	0.687	0.749	0.074	0.103	0.135
First-stage Take-up Rate (LFC for Uniform Sign Tests)											
0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.003	0.003	0.006
Exam-taking Rate (GMS for Uniform Sign Tests)											
-	-	-	0.012	0.007	0.005	0.092	0.099	0.097	-	-	-
Exam Grade (GMS for Uniform Sign Tests)											
-	-	-	0.002	0.002	0.001	0.545	0.579	0.663	-	-	-
First-stage Take-up Rate (GMS for Uniform Sign Tests)											
-	-	-	0.000	0.000	0.000	1.000	1.000	1.000	-	-	-

Notes: Data are from Pop-Eleches and Urquiola (2013). Nonparametric local linear estimations are conducted using the triangular kernel and the undersmoothed CCT bandwidth defined in the simulation section. All simulated critical values are calculated with 1,000 bootstrap repetitions.

7 Conclusion

In this paper, we propose uniform tests for treatment effect heterogeneity under both sharp and fuzzy RD designs. Compared with other methods currently adopted in applied RD studies, our tests have the advantage of being both fully nonparametric and robust to weak identification. Monte Carlo simulations show that our tests have very good small sample performance. We apply our methods to a dataset from Romania and reveal strong evidence of treatment effect heterogeneity previously neglected by the literature.

There are several interesting directions for future research. For example, the proposed testing procedure could be extended to examine how quantile or distributional treatment effects (e.g., Bitler et al., 2008; Shen and Zhang, 2016; Shen, 2017) vary with observed individuals characteristics. The proposed tests could also be extended to examine heterogeneity in rank shuffling (Dong and Shen, 2018), or the marginal threshold treatment effect (Dong and Lewbel, 2015) under the RD or the traditional treatment effect setups.

References

Anderson, M. L. (2008): "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103, 1481–1495.

Andrews, D. W. K. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, ed. by R. F. Engle and D. L. McFadden, North Holland, vol. 4, 2111–3155.

Andrews, D. W. K. 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.

Andrews, D. W. K. and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

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.

Aradillas-Lopez, A., A. Gandhi, and D. Quint (2016): "A Simple Test for Moment Inequality Models with an Application to English Auctions," *Journal of Econometrics*, 194.1, 96–115.

Barrett, G. F. and S. G. Donald (2003): "Consistent Tests for Stochastic Dominance," *Econometrica*, 71, 71–104.

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

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

Bierens, H. J. (1982): "Consistent Model Specification Tests," *Journal of Econometrics*, 20.1, 105–134.

——— (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58.6, 1443–1458.

Bierens, H. J. and W. Ploberger (1997): "Asymptotic Theory of Integrated Conditional Moment Tests," *Econometrica*, 1129–1151.

Bitler, M. P., J. B. Gelbach, and H. W. Hoynes (2008): "Distributional Impacts of the Self-Sufficiency Project," *Journal of Public Economics*, 92, 748–765.

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. (2016): "Regression Discontinuity Designs Using Covariates," *working paper, University of Michigan*.

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

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.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737.

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

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

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.

Dong, Y. and S. Shen (2018): "Testing for Rank Invariance or Similarity in Program Evaluation." *Review of Economics and Statistics*, 100, 78–85.

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 (2015): "Weak Identification in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, forthcoming.

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

Hansen, P. R. (2005): "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23, 365–380.

Hsu, Y. (2017): "Consistent Tests for Conditional Treatment Effects," *Econometric Journal*, 20.1, 1–22.

Hsu, Y.-C. (2016): "Multiplier Bootstrap," Tech. rep., Academia Sinica.

Imbens, G. W. and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615 – 635.

Kosorok, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*, Springer: New York.

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

Lee, S., K. Song, and Y.-J. Whang (2013): "Testing Functional Inequalities," *Journal of Econometrics*, 172.1, 14–32.

——— (2017): "Testing for a General Class of Functional Inequalities," *Econometric Theory*, forthcoming.

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.

Romano, J. P. and A. M. Shaikh (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211.

Shen, S. (2017): "Estimation and Inference of Distributional Partial Effects: Theory and Application," *Journal of Business and Economic Statistics*, forthcoming.

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

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.

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.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer, New York.

Whang, Y.-J. (2000): "Consistent Bootstrap Tests of Parametric Regression Functions," *Journal of Econometrics*, 98.1, 27–46.

——— (2001): "Consistent Specification Testing for Conditional Moment Restrictions," *Economics Letters*, 71.3, 299–306.