2016/26

Regression Discontinuity Design with Many Thresholds

Marinho Bertanha

50 YEARS OF

CORE

DISCUSSION PAPERS

CORE
Voie du Roman Pays 34, L1.03.01
Tel (32 10) 47 43 04
Fax (32 10) 47 43 01
Email: immaq-library@uclouvain.be
http://www.uclouvain.be/en-44508.html

Regression Discontinuity Design with Many Thresholds*

Marinho Bertanha†

April 7, 2016

Abstract

Numerous empirical studies employ regression discontinuity designs with multiple cutoffs and heterogeneous treatments. A common practice is to normalize all the cutoffs to zero and estimate one effect. This procedure identifies the average treatment effect (ATE) on the observed distribution of individuals local to existing cutoffs. However, researchers often want to make inferences on more meaningful ATEs computed over general counterfactual distributions of individuals rather than simply the observed distribution of individuals local to existing cutoffs. This paper proposes a root-n consistent and asymptotically normal estimator for such ATEs when heterogeneity follows a non-parametric function of cutoff characteristics in the sharp case. It shows that identification in the fuzzy case with multiple cutoffs is impossible unless heterogeneity follows a finite dimensional function of cutoff characteristics. Under parametric heterogeneity, this paper proposes an ATE estimator for the fuzzy case that optimally combines observations to minimize its mean squared error.

Keywords: Regression Discontinuity Designs, Multiple Cutoffs, Average Treatment Effect, Alternative Asymptotics, Peer-effects

JEL Classification: C14, C21, C52, I21.

*This paper is based on Chapter 1 of my doctoral dissertation at Stanford University. I'm indebted to Han Hong, Caroline Hoxby, and Guido Imbens for invaluable advice. The paper also benefited from feedback received from seminar participants at Stanford, California Econometrics Conference, Boston University, Cambridge, Iowa, Notre Dame, UcLouvain, UCSD, UCDavis, FGV-EESP, FGV-EPGE, Insper, and PUC-Rio. I thank Tim Bresnahan, Arun Chandrasekhar, Michael Dinerstein, Ivan Fernandez-Val, Ivan Korolev, Michael Leung, Huiyu Li, Jessie Li, Petra Moser, Stephen Terry, and Xiaowei Yu for suggestions and fruitful discussions. I gratefully acknowledge the financial support received as a B.F. Haley and E.S. Shaw Fellow from the Stanford Institute for Economic Policy Research.

†Center for Operations Research and Econometrics - Université catholique de Louvain, Voie du Roman Pays 34, 1348 Louvain-la-Neuve, Belgium. Email: marinho.bertanha@uclouvain.be. Website: www.uclouvain.be/marinho.bertanha.

1 Introduction

Applications of regression discontinuity design (RDD) have become increasingly popular in economics since the late 1990s (Black (1999), Angrist and Lavy (1999), Van der Klaauw (2002)). One of RDD's main advantages is the identification of a local causal effect under minimal functional form assumptions. More recently, with the increasing availability of richer data sets, there have been many applications with multiple cutoffs and treatments (for example, Black, Galdo, and Smith (2007), Egger and Koethenbuerger (2010), De La Mata (2012), Pop-Eleches and Urquiola (2013)). Existing one-cutoff RDD methods applied to each individual cutoff produce many local effects that are estimated using only a few observations near each cutoff. Researchers often prefer one takeaway summary effect that is more precisely estimated by using all the data. The meaning of a summary effect depends crucially on the heterogeneity assumptions and weights imposed on the different local effects.

Many existing applied studies with multiple cutoffs simply normalize all cutoffs to zero and use the one-cutoff estimator. This normalization procedure estimates an average of local treatment effects weighted by the relative density of individuals near each of the cutoffs (Cattaneo, Keele, Titiunik, and Vazquez-Bare (2016)). Such an average effect would be a meaningful summary measure only in two cases: (i) local treatment effects are all identical and the weighting scheme does not matter; or (ii) local treatment effects are heterogeneous but the researcher is only interested in the average effect on the individuals near the existing cutoffs. However, researchers are often interested in combining observed data with assumptions weaker than (i) to make inferences on counterfactual scenarios more general than (ii).[1]

This paper proposes an estimation procedure for an average treatment effect (ATE) that is a more valuable summary measure than the average effect estimated by the normalization procedure for two reasons. First, the researcher explicitly chooses the counterfactual distribution of the ATE, and this distribution may include individuals at or in between existing cutoffs. Second, the researcher does not need to assume any specific functional form for the heterogeneity of treatment effects across different

[1] In a RDD setting with multiple cutoffs and treatments, it is unreasonable to expect that different local treatment effects are always identical. For example, Pop-Eleches and Urquiola (2013) find that the impact of going to a better high school on academic achievement is heterogeneous across students with different ability levels. Another example is De La Mata (2012) who finds that the eligibility for Medicaid benefits, which depends on your income being below a threshold, decreases the probability of having private health insurance more strongly for lower income thresholds.

cutoffs. As example of an application, suppose you are interested in estimating the effect of Medicaid benefits on health care utilization, and you know that Medicaid eligibility is triggered by income cutoffs that varies across states. Existing one-cutoff RDD methods identify the average effect on individuals with income equal to the income cutoffs, but most interesting policy questions require the average effect over the entire range of income values in your data.

The framework for RDD with many thresholds of this paper is introduced using a simple example based on Pop-Eleches and Urquiola (2013), PEU from now on. Using a wealth of variation of nearly 2,000 cutoffs from the high school assignment in Romania, PEU provides rigorous evidence of the impacts of going to a better school on the academic performance of students. The economic logic of their application is briefly summarized as follows. There is a central planner who assigns students to high schools based on students' scores from a placement test. High schools have limited capacities and are ranked by their qualities. The central planner ranks students by their scores and assigns each of them to the best school available. Each student i submits her score X_i (forcing variable) to the central planner who based on the entire distribution of scores determines a minimum test score c_j (cutoff) for admission to each high school j. The quality of high school j is denoted d_j (treatment dose).

The RDD assignment is assumed sharp for now, that is, students attend the best high school available to them based on their score and the cutoffs that apply to them. As the test score crosses an admission threshold c_j, the quality of the school the student attends changes from d_{j-1} to d_j. Local average effects are denoted by $\mathbb{E}[Y_i(d_j) - Y_i(d_{j-1})|X_i = c_j] \equiv \beta(c_j, d_{j-1}, d_j)$, where $Y_i(d)$ is the potential academic achievement student i has if attending a high school of quality d, and $\beta(c, d, d')$ is the treatment effect function. The two sources of heterogeneity for local treatment effects are the different cutoff values and changes in treatment doses across the different cutoffs. PEU is a particularly illustrative application because it exhibits sufficient variation in cutoff and treatment doses to generate ATEs with substantially greater economic relevance than the typical average based on normalizing all of the cutoffs to zero.

There are numerous other examples of RDD with multiple cutoffs and treatments in different fields of economics. For instance, Egger and Koethenbuerger (2010) study the effect of the size of city government councils on municipal expenditures, where council size is determined by population cutoffs. De La Mata (2012) estimates the

effects of Medicaid benefits on healthcare utilization, where Medicaid eligibility is triggered by income cutoffs that vary across states. Agarwal, Chomsisengphet, Mahoney, and Stroebel (2016) and De Giorgi, Drenik, and Seira (2015) look at multiple cutoffs on credit scores that are used by banks to make credit decisions. There is also a variety of applications in education economics. Angrist and Lavy (1999) and Hoxby (2000) use class size rules to estimate the impact of class size on student achievement. In Hoxby (2000), the variation in cutoff values arises from specific school district class size rules. Several studies exploit different school starting dates to estimate the impact of educational attainment on various outcomes, for example, Dobkin and Ferreira (2010), McCrary and Royer (2011). Duflo, Dupas, and Kremer (2011) analyzes school cohorts that are split into low and high-achieving classes based on test scores, where each school has its own cutoff score. Garibaldi, Giavazzi, Ichino, and Rettore (2012) looks at different income cutoffs that determine tuition subsidies to study the impact of tuition payment on the probability of late graduation from university. In short, there are many applications with variation in cutoffs and treatment doses, but a lack of theoretical investigation on how to combine observations from all cutoffs to estimate economically relevant average effects.

The ability to combine different local effects into an average effect depends on how comparable the researcher believes these effects are. The comparability of local treatment effects essentially depends on the heterogeneity of treatment doses and on the heterogeneity of the treatment effect function $\beta(c, d, d')$. This paper considers two types of assumptions regarding these two heterogeneity aspects. The first heterogeneity assumption says that treatment doses are credibly quantifiable by some variable d. For example, PEU find behavioral evidence that average student performance at each school is a good summary measure for school quality. Another example is the case of a single treatment being triggered by varying cutoffs, like in De La Mata (2012) where each state has its own income threshold for Medicaid coverage. The second heterogeneity assumption specifies a parametric functional form for $\beta(c, d, d')$ guided by Economic theory or a priori knowledge of the researcher. For example, in a class size application like Hoxby (2000), one may derive a functional form based on Lazear (2001)'s model of achievement as a function of class size. Another example is Bajari, Hong, Park, and Town (2011) who present a principal-agent model to study how insurers reimburse hospitals, and the marginal reimbursement rate is discontinuous on health expenditures.

This paper proposes a consistent and asymptotically normal estimator for the Average Treatment Effect (ATE) of a counterfactual distribution of treatment assignment specified by the researcher. A counterfactual policy scenario specifies the distribution of (c, d, d'), and the ATE is the integral of $\beta(c, d, d')$ weighted by such distribution. The first heterogeneity assumption allows the researcher to choose counterfactual distributions with support more general than the finite set of cutoff-dose values observed in the data. Existing RDD methods estimate $\beta(c, d, d')$ at the cutoff-dose values given in the data. The estimator proposed in this paper approximates the ATE integral by averaging estimates of $\beta(c, d, d')$ at existing cutoffs using a proper weighting scheme. Under the first heterogeneity assumption with $\beta(c, d, d')$ nonparametric, the proposed ATE estimator is shown to be asymptotically normal at root-n. This requires an asymptotic sequence where both the number of observations and cutoffs grow to infinity, and sufficient conditions on their rate of growth are given.

Many applications of RDD with multiple cutoffs are in fact fuzzy rather than sharp. In the high school assignment example, a student may choose to attend a high school different than the school she is originally eligible to attend. Multiple treatments result in multiple compliance behaviors, and one-cutoff identification results do not apply in the multiple cutoff case. Building on classic definitions of compliance behaviors (Imbens and Rubin (1997)), compliance groups are defined in terms of *changes* in treatment eligibility and receipt. 'Ever-compliers' are those whose treatment received changes if and only if it changes to the treatment dose they become eligible for. It is assumed that individuals never change into a treatment dose different than the dose of eligibility, a 'no-defiance' condition. In terms of the high school example, if the test score of a student currently in school B increases so as to grant her access to school A, no-defiance implies she either chooses to attend school A or stay at school B, and that she is not triggered to attend some other school C.

This paper shows that identification in fuzzy RDD with multiple treatments is impossible unless the class of treatment effect functions of ever-compliers is restricted to a finite dimensional class. There are important empirical analyses of fuzzy RDD with multiple treatments, for example, Angrist and Lavy (1999), Chen and Van der Klaauw (2008), and Hoekstra (2009); nevertheless, this is the first paper to define compliance and study identification in a general causal framework for multi-cutoff fuzzy RDD. Theses results clarifies the interpretation of two-stage least squares (2SLS) estimates in applications of multi-cutoff fuzzy RDD, a common practice in applied work. The

second heterogeneity assumption states the treatment effect function is of a parametric class which allows for consistent and asymptotically normal estimation of ATEs on ever-compliers. Moreover, both in the sharp and fuzzy cases, the second heterogeneity assumption results in efficiency gains because observations from the various cutoffs are optimally combined to minimize the mean squared error (MSE) of the ATE estimator.

The rapid growth in the number of applications of RDD in Economics in the late 1990s was accompanied by substantial theoretical contributions for inference in the one-cutoff case. Identification and estimation in the sharp and fuzzy cases were formalized by Hahn, Todd, and Van der Klaauw (2001); Fan and Gijbels (1996) and Porter (2003) demonstrated low order bias and rate optimality of the local polynomial estimator. Recent theoretical contributions have addressed the optimal bandwidth choice (Imbens and Kalyanaraman (2012)); alternative asymptotic approximations with better finite sample properties (Calonico, Cattaneo, and Titiunik (2014)); quantile treatment effects (Frandsen, Frölich, and Melly (2012)) and kink treatment effects (Dong (2016b)).

More closely related to the contribution of this paper is the study of treatment effect extrapolation in Angrist (2004), Bertanha and Imbens (2014), Dong and Lewbel (2015), Angrist and Rokkanen (2015), and Rokkanen (2015). These last two papers use observations on additional covariates and restrict the relationship between the heterogeneity of treatment effects and these covariates to obtain identification away from the cutoff. Results in this paper differ from their results because the variation of multiple cutoffs and doses identify ATEs over distributions of individuals in between and at cutoffs without restricting the heterogeneity of treatment effects.

The remainder of this paper is organized as follows. Section 2 sets up the notation and generalize existing RDD results to the multiple cutoff case. Section 3 describes the ATE estimator and gives sufficient conditions for asymptotic normality under the first heterogeneity assumption. Section 4 shows that the class of treatment effect functions needs to be restricted to a finite dimensional class for identification of treatment effects in the fuzzy RDD case. Under the first and second heterogeneity assumptions, Section 4 describes the ATE estimator, derives its asymptotic normality, and gives the optimal weighting scheme to minimize the MSE of estimation. All the proofs are found in Appendix A. A Supplemental Appendix B contains auxiliary results used

in this paper.[2]

2 Setup

This section sets up the framework for RDD with multiple cutoffs. The sharp RDD case is assumed, and existing results in the one-cutoff case are generalized to the multiple cutoff case. There are multiple cutoffs c defined on one scalar forcing variable X that assign individuals to different treatment doses. Each treatment dose is denoted by the variable d, but no further assumption is made on the type of variable d at this point. The forcing variable of individual 'i' is denoted by X_i and lives in a compact interval $\mathcal{X} = [\underline{\mathcal{X}}, \overline{\mathcal{X}}]$. The set of possible treatment doses is defined as $\mathcal{D}$. Cutoffs are indexed by $j = 1, \ldots, K$. The cutoffs are ordered such that $c_1 < c_2 < \ldots < c_K$. Sharp RDD assignment means that an individual with forcing variable X_i is deterministically assigned to treatment dose $D_i = D(X_i)$ for some known assignment mapping $D : \mathcal{X} \to \mathcal{D}$

$$D(x) = \begin{cases} d_0 & \text{if } c_0 \leq x < c_1 \\ d_1 & \text{if } c_1 \leq x < c_2 \\ \vdots & \\ d_K & \text{if } c_K \leq x \leq c_{K+1} \end{cases} \tag{1}$$

where $c_0 = \underline{\mathcal{X}}$, $c_{K+1} = \overline{\mathcal{X}}$. Section 4 looks at the Fuzzy RDD case. Each cutoff is characterized by three variables, $\mathbf{c}_j = (c_j, d_{j-1}, d_j)$: the scalar threshold c_j; the treatment dose d_{j-1} the individual receives if $c_{j-1} \leq X_i < c_j$; and the treatment dose d_j the individual receives if $c_j \leq X_i < c_{j+1}$. The schedule of cutoffs and treatment doses is given by the non-random set $\mathcal{C}_K = \{\mathbf{c}_j\}_{j=1}^K$, $\mathcal{C}_K \subseteq \mathcal{C}$, where $\mathcal{C}$ is a subset of $\mathcal{X} \times \mathcal{D} \times \mathcal{D}$ of all possible cutoff and dose values.[3]

Following Rubin's model of potential outcomes (Rubin (1974), Imbens and Lemieux (2008)), the potential outcome for individual 'i' if she receives treatment dose 'd' is denoted as the random variable $Y_i(d)$. The data generating process is summarized as follows. Values for the forcing variable X_i, and potential outcomes $\{Y_i(d)\}_{d \in \mathcal{D}}$ are drawn iid $i = 1, ..., n$ from a joint distribution. Given $D(x)$, these n individuals are

[2]The Supplemental Appendix is available online at *www.uclouvain.be/marinho.bertanha*.

[3]The validity of the RDD depends crucially on the exogeneity of cutoffs and no manipulation of the forcing variable X by individuals. See McCrary (2008) for a test of forcing variable manipulation. Bajari, Hong, Park, and Town (2011) presents a modified RDD estimator that is consistent under forcing variable manipulation in a class of structural models.

assigned to different treatment doses $D_i = D(X_i)$. The observed outcome Y_i is given by

$$Y_i = \sum_{j=0}^{K} Y_i(d_j)\mathbb{I}\{c_j \leq x < c_{j+1}\}$$

The econometrician observes the schedule of cutoffs and treatment doses $D(x)$ and (Y_i, X_i, D_i) for $i = 1, ..., n$. An individual treatment effect $Y_i(d') - Y_i(d)$ is the change in potential outcome caused by a change in treatment dose $d \to d'$, but the focus is on averages of the distribution of treatment effects. The average treatment effect local to a cutoff $\mathbf{c} = (c, d, d')$ is denoted as

$$\beta(\mathbf{c}) \equiv \mathbb{E}[Y_i(d') - Y_i(d)|X_i = c]$$

Continuity of the conditional mean of potential outcomes leads to identification of treatment effects at the cutoff values $\mathbf{c} \in \mathcal{C}_K$. This result is proved in Hahn, Todd, and Van der Klaauw (2001), and it is restated below for the multiple cutoff case.

Lemma 1. *Assume that $\mathbb{E}[Y_i(d)|X_i = x]$ is a continuous function of x for any $d \in \mathcal{D}$. Then, the treatment effect $\beta(\mathbf{c})$ is identified for any $\mathbf{c} = (c, d, d') \in \mathcal{C}_K$:*

$$\beta(\mathbf{c}) = \lim_{e \downarrow 0} \Big\{ \mathbb{E}[Y_i|X_i = c + e] - \mathbb{E}[Y_i|X_i = c - e] \Big\}$$

The goal of this paper is to lay down conditions for combining local average treatment effects across different $\mathbf{c} \in \mathcal{C}_K$ to obtain meaningful average treatment effects. In practice, RDD applications may have one forcing variable but a variety of cutoff-dose values across several sub-populations. In the high-school assignment case, multiple schools in each town-year are observed as well as many town-years. The pooling of cutoff-dose values from many sub-populations produces a much richer set $\mathcal{C}_K$, and it relies on the following assumption.

(Pooling Assumption) *Let $\beta(\mathbf{c}, p)$ denote the local average treatment effect at cutoff $\mathbf{c}$ conditional on individuals being in some sub-population indexed by p. For any $\mathbf{c} = (c, d, d') \in \mathcal{C}$, and any sub-population p, assume $\beta(\mathbf{c}, p) = \beta(\mathbf{c})$.*

The Pooling Assumption does not restrict the distribution of potential outcomes to be the same across different sub-populations. It accommodates common specifications for pooling data in applied work, for example, time-trends and sub-population fixed effects. The Pooling Assumption says that individuals with the same observed forcing variable X_i that undergo the same change in treatment dose $d \to d'$ have the same

average response across different sub-populations. Therefore, results on this paper for multiple cutoffs on one population are immediately applicable to the pooling of several RDD assignments.

Existing data is typically used to estimate an average treatment effect of the observed policy, but estimates may also be used to infer the average effect of new policies. A new policy consists of a distribution of individuals that undergo changes in treatment doses. More generally, for real-valued $\mathbf{c} = (c, d, d')$, each counterfactual policy scenario translates into a cumulative distribution function $F(\mathbf{c})$ for the forcing variable $X = c$ and treatment dose changes (d, d') within the support $\mathcal{C}$. The average treatment effect of the counterfactual policy is the expected value of the treatment effect function $\beta(\mathbf{c})$ under the counterfactual distribution $F(\mathbf{c})$.

$$\mu(F) = \mathbb{E}_F \beta(\mathbf{c}) \tag{2}$$

The parameter $\mu(F)$ defined in Equation (2) summarizes the average effect of a counterfactual policy F if the distribution of unobservables conditional on observables is not affected by the counterfactual policy.

Identification of ATEs over the general support $\mathcal{C}$ requires quantification of treatment doses in a real-valued variable d, and its discussion is deferred to the next sections. This section describes a baseline case covering those applications of RDD where treatment doses are *not* credibly summarized in a real-valued variable d. For example, Hastings, Neilson, and Zimmerman (2013) study the assignment of students into different degree programs in Universities in Chile. There are multiple cutoffs on a test score, but different cutoffs switch students to completely different treatments which are degree programs in Physics, Engineering, Economics, etc. In the baseline case considered in this section, the set of counterfactual distributions is restricted to distributions with discrete support equal to $\mathcal{C}_K$ observed in the data. Existing results for the one-cutoff case are combined to make inferences on averages of local treatment effects where the researcher specifies the weights placed on different cutoffs.

The discrete counterfactual distribution F^{disc} is associated to the probability mass function or weighting scheme $\{\omega_j\}_j$ over the set $\mathcal{C}_K = \{\mathbf{c}_j\}_j$. For example, in the high school application, a new policy may reallocate students with test scores marginally across the existing cutoffs. The weight ω_j represents the probability mass of students with test score equal to c_j that undergo a change in school quality of $d_{j-1} \to d_j$ in the reallocation policy. The parameter of interest is the average effect on these students which is an average of local effects at the existing cutoffs weighted by $\{\omega_j\}_j$

$$\mu(F^{disc}) = \sum_{j=1}^{K} \omega_j \ \beta(\mathbf{c}_j)$$

where by Lemma 1

$$\beta(\mathbf{c}_j) = B_j = \lim_{e \downarrow 0} \{\mathbb{E}[Y_i | X_i = c_j + e] - \mathbb{E}[Y_i | X_i = c_j - e]\} \tag{3}$$

and $\mu(F^{disc})$ is identified for any choice of F^{disc} with support $\mathcal{C}_K$. Estimation is conducted in two steps.

The first step uses observations near each of the cutoffs c_j to non-parametrically estimate B_j using local polynomial regression (LPR). The researcher chooses a bandwidth parameter $h_1 > 0$, a kernel density function $k(.)$, and the order of the polynomial regression $\rho_1 \in \mathbb{Z}_+$.[4] A polynomial in X is fitted on each side of the cutoff, and the estimator $\hat{B}_j$ is the difference between the intercept of these two polynomial regressions.

$$\hat{B}_j = \hat{a}_j^+ - \hat{a}_j^- \tag{4}$$

$$(\hat{a}_j^+, \hat{\mathbf{b}}_j^+) = \underset{(a,\mathbf{b})}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} \left[Y_i - a - b_1(X_i - c_j) - \ldots - b_{\rho_1}(X_i - c_j)^{\rho_1}\right]^2 \tag{5}$$

$$(\hat{a}_j^-, \hat{\mathbf{b}}_j^-) = \underset{(a,\mathbf{b})}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j-} \left[Y_i - a - b_1(X_i - c_j) - \ldots - b_{\rho_1}(X_i - c_j)^{\rho_1}\right]^2 \tag{6}$$

where $v_i^{j+} = \mathbb{I}\{c_j \leq X_i < c_j + h_1\}$, $v_i^{j-} = \mathbb{I}\{c_j - h_1 < X_i < c_j\}$.

In the second step, the researcher averages out $\hat{B}_j$ to obtain the estimator $\hat{\mu}(F^{disc})$.

$$\hat{\mu}(F^{disc}) = \sum_{j=1}^{K} \omega_j \hat{B}_j$$

For the case of one cutoff, the asymptotic distribution of the LPR estimator $\hat{B}_j$ has been derived before by Hahn, Todd, and Van der Klaauw (2001) and Porter (2003) under $n \to \infty$ and $h_1 \to 0$. For a fixed number of cutoffs K, $h_1 \to 0$ implies

[4] Common choices in the applied literature for these are the edge kernel $k(u) = \mathbb{I}\{|u| \leq 1\}(1-u)$, $\rho_1 = 1$ (local linear regression), and the optimal bandwidth proposed by Imbens and Kalyanaraman (2012).

the neighborhoods around each cutoff don't overlap after large n, which makes $\hat{B}_j$ independent of $\hat{B}_l$ for $j \neq l$. Therefore, the asymptotic distribution of $\hat{\mu}(F^{disc})$ is the weighted sum of the asymptotic distributions of each $\hat{B}_j$. Below, the sufficient conditions for the asymptotic normality result in Theorem 1 are listed. The proof of this and all other results in the paper are in Appendix A.

Assumption 1. *The kernel density function $k : \mathbb{R} \to \mathbb{R}$ is symmetric, bounded, has compact support $[-M, M]$, and can be written as the difference of two weakly increasing functions on $\mathbb{R}$.*

Assumption 2. ***(a)*** *The distribution of X_i has probability density function $f(x)$ that is continuous and has bounded support $\mathcal{X} = [\underline{\mathcal{X}}, \overline{\mathcal{X}}]$;* ***(b)*** *$f(x)$ is one time differentiable with bounded derivative $\nabla_x f(x)$.*

Assumption 3. *Let $\rho_1 \in \mathbb{Z}_+$ be the order of the first step LPR, and $d \in \mathcal{D}$ be arbitrary.* ***(a)*** *$R(x, d) = \mathbb{E}[Y_i(d)|X_i = x]$ is $\rho_1 + 1$ times continuously differentiable w.r.t. x with ρ_1+1-th partial derivative $\nabla_x^{\rho_1+1} R(x, d)$;* ***(b)*** *$\sigma^2(x, d) = \mathbb{V}[Y_i(d)|X_i = x]$ is one time continuously differentiable w.r.t. x with partial derivative $\nabla_x \sigma^2(x, d)$, where '$\mathbb{V}$' denotes the variance operator.*

Theorem 1. *Suppose Assumptions 1-3 hold. As $n \to \infty$ and $h_1 \to 0$, assume $nh_1 \to \infty$ and $\sqrt{nh_1} h_1^{\rho_1+1} \to C \in [0, \infty)$. Then,*

$$\sqrt{nh_1}\left(\hat{\mu}(F^{disc}) - \mu(F^{disc})\right) \xrightarrow{d} N\left(C \sum_j \omega_j \mathcal{B}_j;\ \sum_j \omega_j^2 \mathcal{V}_j\right)$$

where

$$\mathcal{B}_j = \frac{1}{(\rho_1+1)!}\left[\nabla_x^{\rho_1+1} m(c_j^+) - (-1)^{\rho_1+1} \nabla_x^{\rho_1+1} m(c_j^-)\right] e_1' \Gamma^{-1} \gamma^* \tag{7}$$

$$\mathcal{V}_j = \frac{\zeta^2(c_j^+) + \zeta^2(c_j^-)}{f(c_j)} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1 \tag{8}$$

and

$$\begin{aligned}
&\nabla_x^{\rho_1+1} m(c_j^+) = \lim_{x \downarrow c_j} \nabla_x^{\rho_1+1} \mathbb{E}[Y_i|X_i = x] = \nabla_x^{\rho_1+1} R(c_j, d_j) \\
&\nabla_x^{\rho_1+1} m(c_j^-) = \lim_{x \uparrow c_j} \nabla_x^{\rho_1+1} \mathbb{E}[Y_i|X_i = x] = \nabla_x^{\rho_1+1} R(c_j, d_{j-1}) \\
&\zeta^2(c_j^+) = \lim_{x \downarrow c_j} \mathbb{V}[Y_i|\ X_i = x] = \sigma^2(c_j, d_j) \\
&\zeta^2(c_j^-) = \lim_{x \uparrow c_j} \mathbb{V}[Y_i|\ X_i = x] = \sigma^2(c_j, d_{j-1})
\end{aligned}$$

$$\Gamma = \begin{bmatrix} \gamma_0 & \dots & \gamma_{\rho_1} \\ \vdots & \vdots & \vdots \\ \gamma_{\rho_1} & \dots & \gamma_{2\rho_1} \end{bmatrix} \text{ and } \Delta = \begin{bmatrix} \delta_0 & \dots & \delta_{\rho_1} \\ \vdots & \vdots & \vdots \\ \delta_{\rho_1} & \dots & \delta_{2\rho_1} \end{bmatrix} \tag{9}$$

$$\gamma^* = [\gamma_{\rho_1+1} \quad \cdots \quad \gamma_{2\rho_1+1}]'$$

$$e_1 \textit{ is the } (\rho_1 + 1 \times 1) \textit{ vector } e_1 = [1\ 0\ 0\ \cdots\ 0]'$$

$$\gamma_d = \int_0^1 k(u)u^d du \textit{ and } \delta_d = \int_0^1 k(u)^2 u^d du$$

In practice, the bandwidth choice h_1 may differ across cutoffs. All results in this paper hold as long as different bandwidth choices converge to zero at the same rate. To perform inference using this asymptotic result, one needs consistent estimators for the asymptotic bias and variance in Equations 7 and 8. The researcher chooses ρ_1 and the kernel density function $k(.)$ which give γ^*, Γ and Δ; the bandwidth value is used to obtain $\hat{C} = \sqrt{nh_1}h_1^{\rho_1+1}$. It remains to estimate $\nabla_x^{\rho_1+1}m(c_j^{\pm})$, $\zeta^2(c_j^{\pm})$, and $f(c_j)$ which is a straightforward non-parametric problem. For the side derivatives $\nabla_x^{\rho_1+1}m(c_j^{\pm})$, a consistent estimator is obtained from a LPR of order ρ_1+1 (Equations 5 and 6) that uses observations from each side of the cutoff c_j. The estimator is simply the slope coefficient on $(x - c_j)^{\rho_1+1}$. Lemma 3 in the Supplemental Appendix B shows consistency of this estimator. The density $f(c_j)$ is consistently estimated by $\widehat{f}(c_j) = (nh_1)^{-1}\sum_{i=1}^n k\left((X_i - c_j)/h_1\right)$. Porter (2003) suggests estimating $\zeta^2(c_j^{\pm})$ by side-limits of conditional means of squared residuals $\widehat{\varepsilon}_i^2$; that is, $\widehat{\zeta}^2(c_j^{\pm}) = 2(\widehat{f}(c_j)nh_1)^{-1}\sum_{i=1}^n v_i^{j\pm}k\left((X_i - c_j)/h_1\right)\widehat{\varepsilon}_i^2$. To obtain $\widehat{\varepsilon}_i$, first compute the smoothed conditional mean: $\widehat{m}(x) = (\widehat{f}(c_j)nh_1)^{-1}\sum_{i=1}^n k\left((X_i - c_j)/h_1\right)\left(Y_i - \sum_{j=1}^K \mathbb{I}\{c_j \leq x\}\widehat{B}_j\right)$. Finally, $\widehat{\varepsilon}_i = Y_i - \widehat{m}(X_i) - \sum_{j=1}^K \mathbb{I}\{c_j \leq x\}\widehat{B}_j$.

These estimates along with Equations 7, 8 and the definitions of Theorem 1 produce estimates for the asymptotic variance and bias of $\hat{\mu}(F^{disc})$. Alternative approaches to inference include bootstrapping (Hardle and Bowman (1988), Neumann and Polzehl (1998)), the empirical likelihood methods of Otsu, Xu, and Matsushita (2015), or the robust confidence intervals of Calonico, Cattaneo, and Titiunik (2014).

3 Average Treatment Effects in the Sharp Case

This section adds the first heterogeneity assumption to the baseline case of sharp RDD with multiple cutoffs described in the previous section. Treatment doses are assumed to be credibly quantifiable in a real-valued variable d. The treatment effect function $\beta(\mathbf{c})$ is assumed to belong to a non-parametric class of functions like in the previous section. The set of all possible cutoff-doses values is a compact and convex set $\mathcal{C} \subseteq \mathcal{X} \times \mathcal{D} \times \mathcal{D}$. This section covers applications like the high school

assignment in PEU, where the treatment dose is quantified in a quality measure for each school. Examples of measures of school quality include the average test score of peers, the average number of teachers or funding per student. The researcher may impose further heterogeneity restrictions to reduce the dimension of $\beta(\mathbf{c})$. For example, linear returns to school quality implies $\beta(c, d, d') = \beta(c, d' - d)$. Another case is that of a binary treatment for varying values of the cutoff like in De La Mata (2012), where each state has its own income threshold for Medicaid coverage. In the case of binary treatment, the treatment effect function depends only on the cutoff value, that is, $\beta(c, d, d') = \beta(c)$.

Variation in cutoff-dose values along with the first heterogeneity assumption allows for inference on ATEs over the entire support $\mathcal{C}$. For example, the average effect of giving Medicaid benefits to an entire neighborhood of individuals; or the average effect on all the students admitted into a new charter school. This section focus on scalar treatment doses d and continuous counterfactual distributions F^{cont}. Minor changes lead to similar results on multivariate d and discrete or mixed counterfactual distributions. The counterfactual distribution is associated to a continuous probability density function $\omega(\mathbf{c})$ with support equal to $\mathcal{C}$. The ATE is defined as:

$$\mu(F^{cont}) = \int_{\mathcal{C}} \omega(\mathbf{c})\beta(\mathbf{c})\ d(\mathbf{c})$$

An infinite amount of data with a set of cutoff-dose values dense in $\mathcal{C}$ identifies the ATE parameter.

Lemma 2. *Assume that (i) $\beta(\mathbf{c})$ is a continuous function in the compact and convex set $\mathcal{C}$; and that (ii) an infinite amount of data means that there is an countably infinite set of cutoff-doses $\mathcal{C}^*$ that is dense in $\mathcal{C}$ such that $\beta(\mathbf{c})$ is identified for every $\mathbf{c} \in \mathcal{C}^*$. Then, $\mu(F^{cont})$ is identified.*

The parameter $\mu(F^{cont})$ is estimated in two steps. The first step is identical to the procedure described in Equations 4, 5 and 6 of Section 2. That is, LPRs produce estimates $\widehat{B}_j,\ j = 1, \ldots, K$. The second step consists of computing the weighted average of the first step estimates using 'corrected weights' $\{\Delta_j\}_j$ instead of simply weighting by $\{\omega(\mathbf{c}_j)\}_j$. The corrected weight Δ_j is the contribution of estimate $\widehat{B}_j$ to the integral $\int_{\mathcal{C}} \omega(\mathbf{c})\widehat{\beta}(\mathbf{c})\ d(\mathbf{c})$ where $\widehat{\beta}(\mathbf{c})$ is a non-parametric estimate of $\beta(\mathbf{c})$. The estimate $\widehat{\beta}(\mathbf{c})$ is produced by regressing $\widehat{B}_j,\ j = 1, \ldots, K$, on multivariate polynomials of order up to ρ_2 of the cutoff-dose values $\mathbf{c}_j,\ j = 1, \ldots, K$, in the h_2-neighborhood of point $\mathbf{c}$. The choice parameters $\rho_2 \in \mathbb{Z}_+$ and $h_2 > 0$ are specified by the researcher.

$$\widehat{\mu}(F^{cont}) = \sum_{j=1}^{K} \Delta_j \hat{B}_j$$

where

$$\Delta_j = \int_{\mathcal{C}} \omega(\mathbf{c}) \frac{det\left(\mathbf{E}(\mathbf{c})'\mathbf{\Omega}(\mathbf{c};h_2)\mathbf{E}_{\mathbf{0}\leftarrow e_j}(\mathbf{c})\right)}{det\left(\mathbf{E}(\mathbf{c})'\mathbf{\Omega}(\mathbf{c};h_2)\mathbf{E}(\mathbf{c})\right)} \quad d(\mathbf{c}) \tag{10}$$

$$\mathbf{\Omega}(\mathbf{c};h_2) = \mathbf{diag}\left\{\Omega_j(\mathbf{c};h_2)\right\}_{j=1}^{K} \ \ \textit{is a } K \times K \textit{ matrix} \tag{11}$$

$$\Omega_j(\mathbf{c};h_2) = k\left(\frac{c-c_j}{h_2}\right) k\left(\frac{d-d_{j-1}}{h_2}\right) k\left(\frac{d'-d_j}{h_2}\right)$$

$$\mathbf{E}(\mathbf{c}) = [E_1(\mathbf{c}), \ \ldots, \ E_K(\mathbf{c})]' \ \ \textit{is a } K \times J \textit{ matrix} \tag{12}$$

$$E_j(\mathbf{c}) = [\ldots p_{\boldsymbol{\gamma}}(\mathbf{c}-\mathbf{c}_j)\ldots]' \ \ \textit{is a } J \times 1 \textit{ vector with all polynomials of the form}$$

$$p_{\boldsymbol{\gamma}}(\mathbf{c}-\mathbf{c}_j) = (c-c_j)^{\gamma_1}(d-d_{j-1})^{\gamma_2}(d'-d_{j-1})^{\gamma_3}$$

$$\textit{such that } \boldsymbol{\gamma} = (\gamma_1,\gamma_2,\gamma_3) \in \mathbb{Z}_+^3, \sum_i \gamma_i \leq \rho_2, \min\{\gamma_2,\gamma_3\} = 0$$

$$J = 2\binom{\rho_2+2}{2} - (\rho_2+1)$$

$$\textit{and the first element of } E_j(\mathbf{c}) \textit{ is } p_{\mathbf{0}}(\mathbf{c}-\mathbf{c}_j) = 1$$

$$\mathbf{E}_{\mathbf{0}\leftarrow e_j}(\mathbf{c}) \textit{ is a } K \times J \textit{ matrix equal to } \mathbf{E}(\mathbf{c}) \textit{ except for the first column} \tag{13}$$

$$\textit{which is replaced by the } K \times 1 \textit{ vector } e_j$$

$$e_j \textit{ has one in its j-th entry and zero otherwise}$$

The formula for Δ_j is obtained as follows. For each point $\mathbf{c} \in \mathcal{C}$, the estimate $\widehat{\beta}(\mathbf{c})$ is obtained by the following least squares minimization

$$\widehat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta}}{\operatorname{argmin}} \left(\widehat{\mathbf{B}} - \mathbf{E}(\mathbf{c})\boldsymbol{\eta}\right)' \mathbf{\Omega}(\mathbf{c};h_2) \left(\widehat{\mathbf{B}} - \mathbf{E}(\mathbf{c})\boldsymbol{\eta}\right)$$

$$\text{where } \widehat{\mathbf{B}} = \left[\widehat{B}_1, \ \ldots, \ \widehat{B}_K\right]' \text{ a } K \times 1 \text{ vector}$$

$$\widehat{\beta}(\mathbf{c}) = e_1'\widehat{\boldsymbol{\eta}} = \widehat{\eta}_1 \text{ (intercept)}$$

Integrating $\widehat{\beta}(\mathbf{c})$,

$$\int_{\mathcal{C}} \omega(\mathbf{c})\widehat{\beta}(\mathbf{c}) \ \ d\mathbf{c} = \int_{\mathcal{C}} \omega(\mathbf{c}) e_1' \left(\mathbf{E}(\mathbf{c})'\mathbf{\Omega}(\mathbf{c};h_2)\mathbf{E}(\mathbf{c})\right)^{-1} \sum_j \Omega_j(\mathbf{c};h_2)E_j(\mathbf{c})\widehat{B}_j \ \ d(\mathbf{c})$$

$$= \sum_j \int_{\mathcal{C}} \omega(\mathbf{c}) e_1' \left(\mathbf{E}(\mathbf{c})'\mathbf{\Omega}(\mathbf{c};h_2)\mathbf{E}(\mathbf{c})\right)^{-1} \Omega_j(\mathbf{c};h_2)E_j(\mathbf{c}) \ \ d(\mathbf{c}) \ \ \widehat{B}_j$$

$$= \sum_j \int_{\mathcal{C}} \omega(\mathbf{c}) \frac{det\left(\mathbf{E}(\mathbf{c})'\mathbf{\Omega}(\mathbf{c};h_2)\mathbf{E}_{\mathbf{0}\leftarrow e_j}(\mathbf{c})\right)}{det\left(\mathbf{E}(\mathbf{c})'\mathbf{\Omega}(\mathbf{c};h_2)\mathbf{E}(\mathbf{c})\right)} \quad d(\mathbf{c}) \quad \widehat{B}_j$$

$$= \sum_j \Delta_j \widehat{B}_j$$

where the third equality uses the Cramer rule.

A necessary condition for consistency of $\hat{\mu}(F^{cont})$ is that the schedule of cutoff-dose values becomes dense in the set $\mathcal{C}$ as the sample size n increases. Different than other sections of this paper, the asymptotic exercise of this section has the sample size $n \to \infty$ and number of cutoffs $K \to \infty$.[5] Assumption 4 restricts the asymptotic behavior of the triangular array of points $\{\mathbf{c}_j\}_{j=1}^K$.

Assumption 4. ***(a)****The schedule of cutoffs and doses comes from a triangular array of fixed constants that depends on the total number of cutoffs K denoted $\mathcal{C}_K = \{\mathbf{c}_{j,K}\}_{j=1}^K \subset \mathcal{C}$;* ***(b)****given the 1st step estimation bandwidth sequence $h_{1,K}$, the estimation windows at two consecutive cutoffs do not overlap; that is, $\exists M \in (0,\infty)$ such that $\min_{j=1,\ldots,K-1} |c_{j+1,K} - c_{j,K}| > M h_{1,K}$ after large K;* ***(c)****given the 2nd step estimation bandwidth sequence $h_{2,K}$ and polynomial order ρ_2, define, for each K, $\mathbf{E}(\mathbf{c})$, and $\mathbf{\Omega}(\mathbf{c};h_{2,K})$ as in equations 11 - 13. There exists a positive definite $J \times J$ matrix $\mathbf{Q}$ such that*

$$\sup_{\mathbf{c}\in\mathcal{C}} \left\| K h_{2,K}^3 \left[\mathbf{E}(\mathbf{c}/h_{2,K})'\mathbf{\Omega}(\mathbf{c};h_{2,K})\mathbf{E}(\mathbf{c}/h_{2,K})\right]^{-1} - \mathbf{Q} \right\| = o(1)$$

For large K, cutoff-dose values must be uniformly distributed on the domain $\mathcal{C}$ such that $\mathbf{E}(\mathbf{c}/h_2)'\mathbf{\Omega}(\mathbf{x};h_2)\mathbf{E}(\mathbf{c}/h_2)$ is invertible and of magnitude Kh_2^3, that is, K times the volume of every h_2-neighborhood of $\mathbf{c}$, for every $\mathbf{c}$ in $\mathcal{C}$. These conditions are satisfied in a variety of examples of triangular arrays of points that cover $\mathcal{C}$ uniformly well for large K. In the Supplemental Appendix (Section B.3), these conditions are verified for one example of triangular array. Asymptotic normality also relies on additional smoothness conditions on the moments of the data.

Assumption 5. ***(a)****$R(x,d)$ (defined in Assumption 3) is a $\bar{\rho}$ times continuously differentiable function with $\bar{\rho} = \max\{\rho_1 + 2, \rho_2 + 1\}$, where ρ_i is the order of the polynomial approximation in the estimation step $i = 1, 2$; the $\bar{\rho}$-th partial derivative of*

[5] In practice, the number of cutoffs may be large due to the pooling of several sub-populations. Having one sub-population with a growing number of cutoffs instead of a growing number of sub-populations is a more tractable exercise. In light of the Pooling Assumption of Section 2, a cutoff $\mathbf{c}$ in an additional sub-population produces the same average treatment effect $\beta(\mathbf{c})$ as if this were an additional cutoff in the existing sub-population.

$R(x,d)$ with respect to x is denoted $\nabla_x^{\bar{\rho}} R(x,d)$; ***(b)*** *$\nabla_x \sigma^2(x,d)$ (defined in Assumption 3) is a continuous function;* ***(c)*** *$\exists M \in (0,\infty)$ such that $\mathbb{P}[|Y_i(d) - R(x,d)| < M] = 1$ for $\forall (x,d)$.*

Theorem 2 states the rate conditions under which the estimator $\widehat{\mu}(F^{cont})$ has an asymptotic normal distribution. Estimation of the ATE consists of approximating the integral of the treatment effect function by a weighted sum of the values of such function at a finite number of points in its domain. The approximation error converges to zero as the number of points grows large. The function evaluations are the estimated $\widehat{B}_j$ which means that the integral approximation error has to converge to zero fast enough to ensure asymptotic normality.

Theorem 2. *Suppose Assumptions 1-5 hold. As $n \to \infty$, assume that $K \to \infty$, $h_1 \to 0$, and $h_2 \to 0$ such that* ***(i)*** *$\sqrt{Knh_1} h_1^{\rho_1+1} \to C \in [0,\infty)$;* ***(ii)*** *$\sqrt{K} \log n / \sqrt{nh_1} \to 0$ and $Kh_1 = O(1)$;* ***(iii)*** *$\sqrt{Knh_1} h_2^{\rho_2+1} \to 0$ and $1/Kh_2^3 = O(1)$; Then*

$$\sqrt{Knh_1}\Big(\widehat{\mu}(F^{cont}) - \mu(F^{cont})\Big) \xrightarrow{d} N\Big(C \cdot \boldsymbol{\mathcal{B}}^{\infty}; \boldsymbol{\mathcal{V}}^{\infty}\Big)$$

where

$$\boldsymbol{\mathcal{B}}^{\infty} = \frac{(e_1' \Gamma^{-1} \gamma^*)}{(\rho_1+1)!} \int_{\mathcal{C}} \omega(\mathbf{c}) \left[\nabla_x^{(\rho_1+1)} R(c,d') - (-1)^{\rho_1+1} \nabla_x^{(\rho_1+1)} R(c,d)\right] \, d\mathbf{c}$$

$$\boldsymbol{\mathcal{V}}^{\infty} = \left(e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1\right) \left(e_1' \mathbf{Q} \Theta\right) \int_{\mathcal{C}} \omega(\mathbf{c})^2 \frac{\sigma^2(c,d') + \sigma^2(c,d)}{f(c)} \, d\mathbf{c}$$

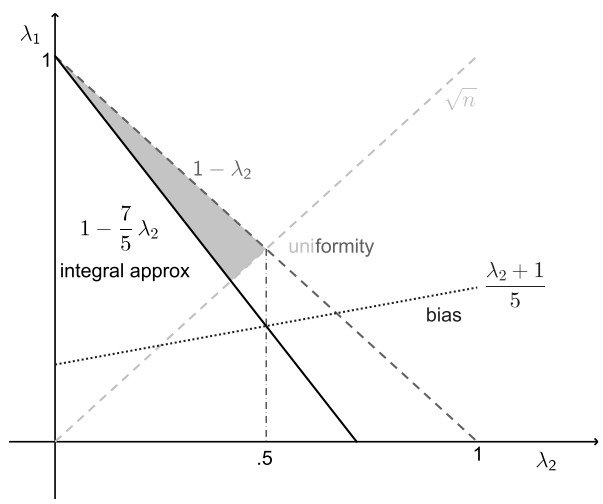
$$\Theta \equiv \lim_{K \to \infty} \int_{\mathcal{C}} \Omega_j(\mathbf{c}; h_2) E_j(\mathbf{c}/h_2) \, d\mathbf{c} \quad \forall j$$

using definitions in Assumptions 2-4, and Equations 9, 11, 12.

A simple example illustrates the rate conditions. Suppose $h_1 = n^{-\lambda_1}$, $K = n^{\lambda_2}$, $\rho_1 = 1$ (local linear regression in the 1st step), $\rho_2 = 3$ (local cubic regression in the 2nd step), and $h_2 = K^{-3/10}$. First, note that $h_2 \to 0$ and $1/Kh_2^3 = O(1)$. The rate conditions on K and h_1 are illustrated in Figure 1 in terms of (λ_1, λ_2). In this setting, the first rate condition gives $\lambda_1 \geq (\lambda_2+1)/(2\rho_1+3) = (\lambda_2+1)/5$: the bandwidth of the first step estimation has to converge to zero fast enough to control the asymptotic bias ('bias', dotted line); the second set of rate conditions gives $\lambda_1 < 1 - \lambda_2$ and $\lambda_2 \leq \lambda_1$: the total number of cutoffs K cannot grow too fast relatively to the sample size n so to have enough observations around the cutoffs to ensure uniform consistency of first step

estimates ('uniformity', dashed lines). The third set of rate conditions is equivalent to $1+\lambda_2\left(1-\frac{6}{10}(\rho_2+1)\right)<\lambda_1$ or $1-\lambda_2\left(\frac{7}{5}\right)<\lambda_1$: K has to grow fast enough relatively to n to ensure the integral approximation error vanishes faster than the variance of the estimator ('integral approx', solid line). Note that ρ_2 must be $\rho_2 \geq 3$ otherwise the 'integral approx' condition $\lambda_1 > 1-\lambda_2\left(\frac{3\rho_2-2}{5}\right)$ contradicts the 'uniformity' condition that says $\lambda_1 < 1-\lambda_2$. The shaded area in Figure 1 below illustrates the set of choices for the bandwidth power λ_1 for a given $\lambda_2 \in (0, .5)$. In this example, there is no asymptotic bias unless $\rho_2 \geq 6$. The maximum convergence rate of the estimator $\sqrt{Knh_1}$ is equal to $\sqrt{n}$ along the dashed line labeled $\sqrt{n}$. Figure 1 illustrates the feasibility set for choosing the bandwidth sequence h_1 given (h_2, K, ρ_1, ρ_2) according to the rate conditions of Theorem 2. The schedule of cutoff-doses has to satisfy Assumption 4 for (h_1, h_2, ρ_2) in such feasibility set. In the Supplemental Appendix (Section B.3) these conditions are verified for one example of triangular arrays.

Figure 1: Rate Conditions of Theorem 2



Notes: Rate conditions of Theorem 2 for the example of sequences $h_1 = n^{-\lambda_1}$, $K = n^{\lambda_2}$, and $h_2 = K^{-3/10}$. The first step estimation uses $\rho_1 = 1$ (local linear regression), and the second step estimation uses $\rho_2 = 3$ (local cubic regression). The rate conditions are depicted as following: $\sqrt{Knh_1}h_1^{\rho_1+1} \to C$ ('bias', dotted line); $\sqrt{K}\log n/\sqrt{nh_1} \to 0$ and $Kh_1 = O(1)$ ('uniformity', dashed lines); and $\sqrt{Knh_1}{h_2}^{\rho_2+1} \to 0$ ('integral approx', solid line).

The asymptotic bias and variance terms of Theorem 2 are consistently estimated

using the procedures discussed at the end of Section 2. The proof of Theorem 2 shows that $\boldsymbol{\mathcal{B}}^{\infty} = \lim_{K\to\infty} \sum_j \Delta_j \mathcal{B}_j$ and $\boldsymbol{\mathcal{V}}^{\infty} = \lim_{K\to\infty} K \sum_j \Delta_j^2 \mathcal{V}_j$ where $\mathcal{B}_j$, $\mathcal{V}_j$, and Δ_j are defined in Equations 7, 8, 10. Once $\widehat{\mathcal{B}}_j$, $\widehat{\mathcal{V}}_j$, and Δ_j are computed, the asymptotic bias and variance terms are estimated by computing the weighted sums above. Sufficient moment and rate conditions give consistency of these estimators under the asymptotics with large number of cutoffs.

4 Fuzzy Case with Multiple Cutoffs

This section relaxes the sharp assignment mechanism of previous sections. Identification requires further restrictions on the treatment effect function, and a second heterogeneity assumption is imposed. Namely, the treatment effect function is assumed to be in a parametric class of functions. As a result, asymptotic inference does not require a large number of cutoffs but only a large sample size.

In the sharp RDD case, all individuals with forcing variable equal to x receive the same treatment $D(x)$ (defined in Eq. 1). In the fuzzy RDD case, many of these individuals may receive different treatments for unobserved reasons. In the high school assignment example, students may choose to go to a school that is not the best school they get in. For instance, a student may want to attend the same high school as does a certain friend or sibling. Another example is Garibaldi, Giavazzi, Ichino, and Rettore (2012), where the schedule of tuition subsidies applies to most students in Bocconi University, but the university reserves the right to grant certain students different subsidies after reassessing their ability to pay.[6]

The fuzzy RDD case is modeled in terms of the potential treatment assignment framework. Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote a probability space for the population of interest. For each individual i, the potential treatment outcome is described by the measurable function $\mathcal{U}_i : \mathcal{X} \to \mathcal{D}$. The function $\mathcal{U}_i(x)$ describes the treatment individual i would be assigned to in case her forcing variable were equal to x. For simplicity, it is assumed to belong to the following class of functions

[6]The source of fuzziness varies across applications. One example is the case where the assignment of individuals into different treatments is made through a matching mechanism, and the econometrician does not observe all the individual characteristics used in the matching algorithm. This is the reason why the RDD in PEU is fuzzy: based on the entire distribution of test scores and preferences, the central planner ranks students by their test scores and assigns each one of them to her most preferred school among those schools with vacancies.

$$\mathcal{U}^* = \Bigg\{ \mathcal{U} : \mathcal{X} \to \mathcal{D} \ : \ \mathcal{U}(x) = \sum_{j=0}^{K} u_j \mathbb{I}\left\{c_j \leq x < c_{j+1}\right\}$$

$$\text{for some } u_j \in \{d_0, \ldots, d_K\}, j = 0, \ldots, K \Bigg\} \tag{14}$$

Sharp RDD is the particular case where the potential treatment assignment function is the same for every individual $\mathcal{U}_i(x) = D(x)$. Potential treatment functions $\mathcal{U}_i(x)$ are unobserved, but the treatment received is observed and given by

$$D_i = \sum_{j=0}^{K} \mathcal{U}_i(c_j) \mathbb{I}\left\{c_j \leq X_i < c_{j+1}\right\}$$

Using classic definitions of compliance behaviors (Imbens and Rubin (1997)), three types of compliance groups are defined in terms of *changes* in treatment eligibility. 'Never-changers' are those whose treatment received never changes when eligibility changes. The treatment received by 'ever-compliers' or 'ever-defiers' changes at least once when eligibility changes. 'Ever-compliers' are those whose treatment received changes if and only if it changes to the treatment dose they become eligible for. 'Ever-defiers' change to a treatment dose different from the one they become eligible for. In the case of one cutoff and two treatments, the definition of ever-complier (ever-defier) is equivalent to the classic definition of complier (defier) of Imbens and Lemieux (2008).

The three compliance groups are measurable sets that partition the Ω space with $\mathbf{G}_{nc}$ denoting 'never-changers, $\mathbf{G}_{ec}$ 'ever-compliers', and $\mathbf{G}_{ed}$ 'ever-defiers'.[7]

$$\mathbf{G}_{nc} = \Big\{ \omega \in \Omega : \{j : \ \mathcal{U}_i(c_{j-1}) \neq \mathcal{U}_i(c_j)\} = \varnothing \Big\} \tag{15}$$

$$\mathbf{G}_{ec} = \Big\{ \omega \in \Omega : \{j : \ \mathcal{U}_i(c_{j-1}) \neq \mathcal{U}_i(c_j)\} = \{j : \ \mathcal{U}_i(c_j) = D(c_j)\} \neq \varnothing \Big\} \tag{16}$$

$$\mathbf{G}_{ed} = \Big\{ \omega \in \Omega : \{j : \ \mathcal{U}_i(c_{j-1}) \neq \mathcal{U}_i(c_j)\} \supseteq \{j : \ \mathcal{U}_i(c_j) \neq D(c_j)\} \neq \varnothing \Big\} \tag{17}$$

In the high school assignment case, an example of a 'never-changer' is a student

[7] These definitions allow for non-monotonic treatment schedules. For example, in the class-size rule applications when the class size drops and increases after each cutoff. Table 3 in the Supplemental Appendix B (Section B.4) illustrates these definitions of compliance groups using a simple example with 3 treatments and 2 cutoffs.

who strongly prefers the high school with the lowest admission cutoff and attends that high school even if she is admitted to better schools. An example of an 'ever-complier' is a student who attends the best school into which she is admitted or a student who chooses the best school among those nearby schools. If a student has rational preferences, is never indifferent, and always picks a high school among those schools with admission cutoffs that are less or equal than her test score, then she is never an 'ever-defier'. In other words, as her test score increases, a new school is added to her choice-set of schools; she either chooses to go to the new school to which she becomes eligible for, or she stays at the school which she preferred the most prior to the increase in her choice-set. It is plausible to assume 'no-ever-defiance' for this and other applications.

Never-changers do not produce changes in treatments, so there is no identification on them. For ever-compliers, there are multiple possible changes in treatment at a given cutoff, and ever-compliers may differ in terms of the treatments they comply with. For example, the student who is willing to attend the best school possible complies with all changes in treatment eligibility. On the other hand, the student who is willing to attend the best school possible within a certain distance from home only complies with some of the changes in treatment eligibility. Therefore, besides no-defiance, identification also requires the heterogeneity of ever-compliers to be restricted.

Assumption 6 is a generalized version of the sufficient conditions for identification of the treatment effect on compliers in the case with one cutoff as in Hahn, Todd, and Van der Klaauw (2001) and Dong (2016a). It also states that the functional form of treatment effects does not vary across different types of ever-compliers.

Assumption 6. ***(a)*** *There are no ever-defiers:* $\mathbb{P}[\mathbf{G}_{ed}] = 0$*;* ***(b)*** *for arbitrary* $d \in \mathcal{D}$*, and* $\bar{\mathcal{U}}$ *in the class* $\mathcal{U}^*$ *defined in Eq. 14,* $\mathbb{E}[Y_i(d)|X_i = x, \mathcal{U}_i = \bar{\mathcal{U}}]$ *and* $\mathbb{P}\left[\mathcal{U}_i = \bar{\mathcal{U}}|X_i = x\right]$ *are continuous and bounded functions of* x*;* ***(c)*** *for arbitrary* $\mathbf{c} = (c, d, d') \in \mathcal{C}$*, ever-compliers share the same treatment effect function:* $\beta_{ec}(\mathbf{c}) \equiv \mathbb{E}[Y_i(d') - Y_i(d)|X_i = c, \mathcal{U}_i = \bar{\mathcal{U}}]$ *for all* $\bar{\mathcal{U}} \in \mathcal{U}^*$ *such that* $\left\{\mathcal{U}_i = \bar{\mathcal{U}}\right\} \subseteq \mathbf{G}_{ec}$*, for* $\mathbf{G}_{ec}$ *defined in Eq. 16.*

A fuzzy assignment produces several different treatment changes at each cutoff even after ruling out ever-defiers. The researcher only observes one aggregate change in Y_i at each cutoff, but there are have several possible treatment effects on ever-compliers to be identified. Identification of these effects is not possible without

further restricting the class of functions $\beta_{ec}(\mathbf{c})$ belongs to. Economic theory or a priori knowledge guides the choice of a functional form that credibly summarizes the heterogeneity of treatment effects. For example, the principal-agent model of Bajari, Hong, Park, and Town (2011) is applied to study reimbursement of hospitals by insurers in a setting where the marginal reimbursement rate is discontinuous on health expenditures.

The second heterogeneity assumption restricts the treatment effect function on ever-compliers to a finite dimensional vector space of functions as follows.

Assumption 7. *Let $\boldsymbol{\mathcal{W}}(c,d) = [\mathcal{W}_1(c,d), \ldots, \mathcal{W}_q(c,d)]'$ be a vector valued function $\boldsymbol{\mathcal{W}} : \mathcal{X} \times \mathcal{D} \to \mathbb{R}^{q \times 1}$ known to the researcher and such that **(a)** $\mathbb{E}_F\left[\boldsymbol{\mathcal{W}}(c,d') - \boldsymbol{\mathcal{W}}(c,d)\right]$ is well defined for the counterfactual distribution F; and **(b)** $\mathcal{W}_j(c,d') - \mathcal{W}_j(c,d)$, $j = 1, \ldots, q$, are linearly independent functions; the treatment effect function $\beta_{ec}(\mathbf{c})$ is assumed to belong to the following class of functions*[8]

$$\mathcal{H} = \left\{ \beta : \mathcal{C} \to \mathbb{R} \ : \ \beta(c,d,d') = \left[\boldsymbol{\mathcal{W}}(c,d') - \boldsymbol{\mathcal{W}}(c,d)\right]' \boldsymbol{\theta}, \text{ for } \boldsymbol{\theta} \in \mathbb{R}^q \right\}$$

In this case, the ATE on ever-compliers is a linear combination of the true parameter vector $\boldsymbol{\theta}_0^{ec}$. For a counterfactual distribution F chosen by the researcher

$$\mu_{ec}(F) = \mathbb{E}_F\left[\beta(\mathbf{c}; \boldsymbol{\theta}_0^{ec})\right] = \mathbb{E}_F\left[\boldsymbol{\mathcal{W}}(c,d') - \boldsymbol{\mathcal{W}}(c,d)\right]' \boldsymbol{\theta}_0^{ec} \equiv \mathbf{Z}(F)\boldsymbol{\theta}_0^{ec} \tag{18}$$

where $\mathbf{Z}(F) = \mathbb{E}_F\left[\boldsymbol{\mathcal{W}}(c,d') - \boldsymbol{\mathcal{W}}(c,d)\right]'$ is a known $1 \times q$ vector.

Theorem 3 shows that the observed change in average outcome at a given cutoff is a weighted average of treatment effects on ever-compliers who switch from various doses into the dose associated with being on the right of that cutoff. Assumption 7 and variation in cutoff characteristics are sufficient conditions for identification. Conversely, identification on ever-compliers implies that $\beta_{ec}(\mathbf{c})$ belongs to a finite dimensional class of functions. As in the sharp case, the Pooling Assumption of Section 2 applied to β_{ec} allows one to combine observations from many sub-populations.

[8] Assumption 7 imposes a parametric functional form on $\beta(\mathbf{c})$ which is equivalent to $\mathbb{E}[Y_i(d)|X_i = c]$ having a semi-parametric functional form. This is a weaker restriction than the parametric $\mathbb{E}[Y_i(d)|X_i = c]$ assumed in some empirical analyses of RDD. Empirical evidence suggests the treatment effect function to be a less complex function than the conditional mean of outcomes. In this case, misspecifying $\mathbb{E}[Y_i(d)|X_i = c]$ leads to greater bias than misspecifying $\beta(\mathbf{c})$, and Assumption 7 is preferred to a parametric assumption on $\mathbb{E}[Y_i(d)|X_i = c]$.

Theorem 3. *Under Assumption 6, for $j = 1, \ldots, K$*

$$B_j = \sum_{l=0, l\neq j}^{K} \omega_{j,l} \beta_{ec}(c_j, d_l, d_j)$$

where B_j is defined in Equation 3 and, for $l = 0, 1, \ldots, K, l \neq j$

$$\omega_{j,l} = \lim_{e \downarrow 0} \{\mathbb{P}[D_i = d_l | X_i = c_j - e] - \mathbb{P}[D_i = d_l | X_i = c_j + e]\}$$

Moreover, suppose β_{ec} belongs to the class of functions $\mathcal{H}$ defined in Assumption 7 with $q \leq K$. Define

$$\widetilde{W}_j = \sum_{l=0, l\neq j}^{K} \omega_{j,l} \left[\boldsymbol{\mathcal{W}}(c_j, d_j) - \boldsymbol{\mathcal{W}}(c_j, d_l)\right] \tag{19}$$

for the vector valued function $\boldsymbol{\mathcal{W}}(c, d)$ of Assumption 7; and define the $K \times q$ matrix $\widetilde{\mathbf{W}}$ by stacking $\widetilde{W}_j$, and $\mathbf{B}$ by stacking B_j. If $\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}}$ is invertible, then $\beta_{ec}(\mathbf{c})$ is identified and equal to

$$\beta_{ec}(\mathbf{c}) = [\boldsymbol{\mathcal{W}}(c, d') - \boldsymbol{\mathcal{W}}(c, d)]' \left(\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}}\right)^{-1} \widetilde{\mathbf{W}}'\mathbf{B}.$$

Conversely, suppose β_{ec} belongs to some class of functions $\widetilde{\mathcal{H}}$, and treatment effects on ever-compliers are identified at the $p > K$ cutoff-dose values $\{\tilde{\mathbf{c}} : \tilde{\mathbf{c}} = (c_j, d_l, d_j)$ with $\omega_{j,l} > 0\}$ of every fuzzy assignment generated from the given schedule of cutoffs $\{\mathbf{c}_j\}_{j=1}^{K}$. Then, the class of functions $\widetilde{\mathcal{H}}$ is 'finite dimensional' in the sense that

$$\mathcal{G} = \left\{\left(\beta(\tilde{\mathbf{c}}_1), \ldots, \beta(\tilde{\mathbf{c}}_p)\right) : \text{ for } \beta \in \widetilde{\mathcal{H}}\right\} \subseteq \mathbb{R}^p$$

has $dim\mathcal{G} \leq K$ for every fuzzy assignment $\{\tilde{\mathbf{c}}_j\}_{j=1}^{p}$ generated from $\{\mathbf{c}_j\}_{j=1}^{K}$.

Contrary to the case of a single cutoff, Theorem 3 reveals the requirement of stronger functional form assumptions on $\beta_{ec}(\mathbf{c})$ for identification in the fuzzy case with multiple cutoffs. For example, identification is not possible when $\widetilde{\mathcal{H}}$ is the class of all smooth functions studied in the non-parametric case (Section 3). Theorem 3 also clarifies the interpretation of two-stage least squares (2SLS) estimates in applications of fuzzy RD with multiple cutoffs, a common practice in applied work. The practice consists of instrumenting D_i with $D(X_i)$ in a regression of Y_i on a constant, D_i, and X_i (see Angrist and Pischke (2008)) for a discussion.

In the single cutoff case, both the non-parametric RD estimator and 2SLS applied to a neighborhood of the cutoff are consistent to the average treatment effect on compliers (Hahn, Todd, and Van der Klaauw (2001)). To my knowledge, such an equivalence has never been studied in the multiple cutoff case. Nevertheless, there

are many important applications with fuzzy multiple-cutoff RD that use 2SLS, for example, Angrist and Lavy (1999), Chen and Van der Klaauw (2008), Hoekstra (2009), Pop-Eleches and Urquiola (2013). The 2SLS estimator is consistent for a data-driven weighted average of treatment effects on ever-compliers as long as a sufficiently flexible specification is used (for example, cutoff fixed-effects or varying slopes). The economic meaning of the 2SLS estimands depends crucially on the choice of such weighting scheme. Unless a parametric functional form is imposed on $\beta_{ec}(\mathbf{c})$, only a data-driven weighted average of $\beta_{ec}(\mathbf{c})$ is identified. In other words, if $\beta_{ec}(\mathbf{c})$ is non-parametric, the researcher does not have control over the weighting scheme, and the interpretation of the 2SLS estimands is compromised.

Theorem 3 suggests a two step estimation procedure for $\boldsymbol{\theta}_0^{ec}$, where estimates of B_j are regressed on estimates of $\widetilde{W}_j$. In the first step, in addition to the estimates $\widehat{B}_j$ (Equations 4, 5, and 6), estimates $\widehat{\widetilde{W}}_j$ are obtained using LPRs of $\boldsymbol{\mathcal{W}}(c_j, D_i)$ on X_i at each side of the cutoff c_j. For each $j = 1, \ldots, K$, and l-th coordinate of the vector $\widetilde{W}_j$, $l = 1, \ldots, q$, compute

$$\widehat{\widetilde{W}}_j^{(l)} = \hat{a}_j^{(l)+} - \hat{a}_j^{(l)-}$$

$$(\hat{a}_j^{(l)+}, \hat{\mathbf{b}}_j^{(l)+}) = \underset{(a,\mathbf{b})}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} \left[e_l' \boldsymbol{\mathcal{W}}(c_j, D_i) - a - b_1(X_i - c_j) - \ldots - b_{\rho_1}(X_i - c_j)^{\rho_1}\right]^2$$

$$(\hat{a}_j^{(l)-}, \hat{\mathbf{b}}_j^{(l)-}) = \underset{(a,\mathbf{b})}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j-} \left[e_l' \boldsymbol{\mathcal{W}}(c_j, D_i) - a - b_1(X_i - c_j) - \ldots - b_{\rho_1}(X_i - c_j)^{\rho_1}\right]^2$$

where e_l is the $q \times 1$ vector of zeros except for 1 in its lth coordinate.

The $q \times 1$ vector $\widehat{\widetilde{W}}_j$ is simply $\widehat{\widetilde{W}}_j = \left[\widehat{\widetilde{W}}_j^{(1)}, \ldots, \widehat{\widetilde{W}}_j^{(q)}\right]'$. The regression of $\hat{B}_j$ on $\widehat{\widetilde{W}}_j$ gives an estimate for $\boldsymbol{\theta}_0^{ec}$. More specifically, stack all $q \times 1$ vectors $\widehat{\widetilde{W}}_j$ into the $K \times q$ matrix $\widehat{\widetilde{\mathbf{W}}}$, and $\hat{B}_j$ into the $K \times 1$ vector $\widehat{\mathbf{B}}$. Given the choice of a $K \times K$ symmetric and positive definite weighting matrix Ω, the estimator for $\boldsymbol{\theta}_0^{ec}$ is the solution to the following weighted least squares problem:

$$\widehat{\boldsymbol{\theta}^{ec}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left(\widehat{\mathbf{B}} - \widehat{\widetilde{\mathbf{W}}}\boldsymbol{\theta}\right)' \Omega \left(\widehat{\mathbf{B}} - \widehat{\widetilde{\mathbf{W}}}\boldsymbol{\theta}\right)$$

The estimator for the ATE over ever-compliers $\mu_{ec}(F)$ is a linear combination of

$\widehat{\boldsymbol{\theta}^{ec}}$,

$$\widehat{\mu_{ec}}(F) = \mathbf{Z}(F)\widehat{\boldsymbol{\theta}^{ec}}$$

where $\mathbf{Z}(F)$ is defined in Equation 18.

Asymptotic normality of $\widehat{\boldsymbol{\theta}^{ec}}$ relies on smoothness assumptions on the conditional moments of Y_i and probabilities of treatment for the different compliance behaviors. The sample size grows large while the number of cutoffs remains fixed.

Assumption 8. *For any $d \in \mathcal{D}$, and any $\bar{\mathcal{U}}$ in the class $\mathcal{U}^*$ defined in Eq. 14,* ***(a)***$\mathbb{E}[Y_i(d)|X_i = x, \mathcal{U}_i = \bar{\mathcal{U}}]$ *is $\rho_1 + 1$ times continuously differentiable function of x with $\rho_1 + 1$-th derivative $\nabla_x^{\rho_1+1}\mathbb{E}[Y_i(d)|X_i = x, \mathcal{U}_i = \bar{\mathcal{U}}]$;* ***(b)***$\mathbb{E}[Y_i(d)^2|X_i = x, \mathcal{U}_i = \bar{\mathcal{U}}]$ *is a continuous function of x;* ***(c)****for $\widetilde{\mathbf{W}}$ defined in Eq. 19, $\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}}$ is invertible;* ***(d)***$\mathbb{P}[\mathcal{U}_i = \bar{\mathcal{U}}|X_i = x]$ *is a $\rho_1 + 1$ times continuously differentiable function of x with $\rho_1 + 1$-th derivative $\nabla_x^{\rho_1+1}\mathbb{P}[\mathcal{U}_i = \bar{\mathcal{U}}|X_i = x]$.*

Theorem 4. *Suppose Assumptions 1, 2, 6-8 hold. As $n \to \infty$ and $h_1 \to 0$, assume $nh_1 \to \infty$ and $\sqrt{nh_1}h_1^{\rho_1+1} \to C \in [0, \infty)$. Then,*

$$\sqrt{nh_1}\left(\widehat{\boldsymbol{\theta}}^{ec} - \boldsymbol{\theta}_0^{ec}\right) \xrightarrow{d} N\Bigg(C\left(\widetilde{\mathbf{W}}'\Omega\widetilde{\mathbf{W}}\right)^{-1}\widetilde{\mathbf{W}}'\Omega\boldsymbol{\mathcal{B}}^{ec};$$
$$\left(\widetilde{\mathbf{W}}'\Omega\widetilde{\mathbf{W}}\right)^{-1}\widetilde{\mathbf{W}}'\Omega\boldsymbol{\mathcal{V}}^{ec}\Omega\widetilde{\mathbf{W}}\left(\widetilde{\mathbf{W}}'\Omega\widetilde{\mathbf{W}}\right)^{-1}\Bigg)$$

$$\sqrt{nh_1}\left(\widehat{\mu}_{ec}(F) - \mu_{ec}(F)\right) \xrightarrow{d} N\Bigg(C\mathbf{Z}(F)\left(\widetilde{\mathbf{W}}'\Omega\widetilde{\mathbf{W}}\right)^{-1}\widetilde{\mathbf{W}}'\Omega\boldsymbol{\mathcal{B}}^{ec};$$
$$\mathbf{Z}(F)\left(\widetilde{\mathbf{W}}'\Omega\widetilde{\mathbf{W}}\right)^{-1}\widetilde{\mathbf{W}}'\Omega\boldsymbol{\mathcal{V}}^{ec}\Omega\widetilde{\mathbf{W}}\left(\widetilde{\mathbf{W}}'\Omega\widetilde{\mathbf{W}}\right)^{-1}\mathbf{Z}(F)'\Bigg)$$

where

$\boldsymbol{\mathcal{B}}^{ec} = [\mathcal{B}_1^{ec}, \ldots, \mathcal{B}_K^{ec}]'$, $\mathcal{B}_j^{ec} = [1 \quad -\boldsymbol{\theta}_0^{ec'}]\boldsymbol{\Psi}_j$,

$\boldsymbol{\Psi}_j = \left[\nabla_x^{\rho_1+1}\mathbf{m}(c_j^+) - (-1)^{\rho_1+1}\nabla_x^{\rho_1+1}\mathbf{m}(c_j^-)\right](e_1'\Gamma^{-1}\gamma^*)/(\rho_1+1)!$,

$\mathbf{m}(c_j^{\pm}) = \lim_{x \to c_j^{\pm}} \nabla_x^{\rho_1+1}\mathbb{E}[\mathbf{Y}_{i,j}|X_i = x]$,

$\mathbf{Y}_{i,j} \equiv [Y_i \quad \boldsymbol{\mathcal{W}}(c_j, D_i)]'$,

$\boldsymbol{\mathcal{V}}^{ec} = diag\{\mathcal{V}_1^{ec}, \ldots, \mathcal{V}_K^{ec}\}$, $\mathcal{V}_j^{ec} = [1 \quad -\boldsymbol{\theta}_0^{ec'}]\boldsymbol{\Phi}_j[1 \quad -\boldsymbol{\theta}_0^{ec'}]'$,

$\boldsymbol{\Phi}_j = \left(\boldsymbol{\zeta}^2(c_j^+) + \boldsymbol{\zeta}^2(c_j^-)\right)(e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1)/f(c_j)$,

$\boldsymbol{\zeta}^2(x) = \mathbb{V}[\mathbf{Y}_{i,j}|X_i = x]$,

e_1, Γ, γ^*, Δ *are defined in Equation 9.*

Moreover, the asymptotic mean squared error (MSE) of either $\sqrt{nh_1}\left(\widehat{\boldsymbol{\theta}}^{ec} - \boldsymbol{\theta}_0^{ec}\right)$ or $\sqrt{nh_1}\left(\widehat{\mu}_{ec}(F) - \mu_{ec}(F)\right)$ is minimized when $\Omega = (C^2\boldsymbol{\mathcal{B}}^{ec}\boldsymbol{\mathcal{B}}^{ec'} + \boldsymbol{\mathcal{V}}^{ec})^{-1}$.

The result above along with the MSE-optimal weighting matrix Ω also applies to the Sharp RDD case with parametric treatment effect function. In that case, $\Omega = (C^2 \boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}' + \boldsymbol{\mathcal{V}})^{-1}$ where $\boldsymbol{\mathcal{B}} = [\mathcal{B}_1, \ldots, \mathcal{B}_K]'$, $\boldsymbol{\mathcal{V}} = diag\{\mathcal{V}_1, \ldots, \mathcal{V}_K\}$, and $\mathcal{B}_j$, $\mathcal{V}_j$ are defined in Equations 7 and 8. Moreover, in the Sharp RDD case, root-n normality is obtained under conditions similar to Theorem 2 except for the requirement that K grows fast enough relatively to n.

Estimates for $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Psi}_j$ are computed in a similar fashion as the estimators for $\mathcal{B}_j$ and $\mathcal{V}_j$ described at the end of Section 2. The only difference is that here $\mathbf{Y}_{i,j}$ is $(q+1)$-dimensional while there Y_i is a scalar. The researcher may set Ω equal to the identity matrix to obtain a first round of estimates $\widetilde{\boldsymbol{\theta}}^{ec}$, $\widetilde{\boldsymbol{\mathcal{B}}}^{ec}$, and $\widetilde{\boldsymbol{\mathcal{V}}}^{ec}$. Using $C = \sqrt{nh_1}h_1^{\rho_1+1}$, an estimate for the optimal Ω is obtained by $\widetilde{\Omega} = (C^2 \widetilde{\boldsymbol{\mathcal{B}}}^{ec}\widetilde{\boldsymbol{\mathcal{B}}}^{ec'} + \widetilde{\boldsymbol{\mathcal{V}}}^{ec})^{-1}$. Then, $\widetilde{\Omega}$ may be used to obtain a new estimate $\widehat{\boldsymbol{\theta}}^{ec}$. Similar to the over-identified case of the Generalized Method of Moments, the number of times $\widehat{\boldsymbol{\theta}}^{ec}$ is recomputed does not affect the asymptotic conclusions above.

5 Application

In this section, the methods proposed in this paper are illustrated using the data from PEU on the high school assignment in Romania.[9] Many policy questions demand an ATE of a continuous counterfactual distribution of treatment assignments but the existing one-cutoff RDD methods are not consistent for such ATE. This section provides an example of such policy question. Using the ATE estimator provided in this paper leads to different conclusions than conclusions based on existing one-cutoff RDD techniques. The estimators designed for the sharp RDD case are consistent for 'Intent-to-Treat' (ITT) average effects when applied to the fuzzy data of PEU. In this application, the ITT effect measures the impact of being assigned to a better school but not necessarily attending it. Identification of the effect of attending a better school requires a parametric treatment effect function. The parametric specification yields noticeable efficiency gains in this application and reveals effects on ever-compliers that are much larger than ITT effects.

The administrative data from Romania covers 3 cohorts of 9 grade students for the years of 2001, 2002 and 2003, with a total of 334,137 observations. The essential

[9]The data set is available online in the supplement materials of PEU on the website of the *American Economic Review*.

elements of the high school assignment in Romania are described below, and the reader is referred to PEU for further details. The assignment to high school is nationally centralized by the Ministry of Education. At the end of grade 8, students submit a transition score and a complete ranking of preferences for high schools. The transition score is an average of the student's performance in a national exam taken in grade 8 and the student's grade point average during grades 5-8. The Ministry of Education ranks students by their transition score and no other criteria. The mechanism assigns the student ranked first to her most preferred school, the student ranked second to her most preferred school, etc. Students cannot decline their assignment and have incentives to truthfully reveal their preference rankings.

The observed variables are the town and year of the student i, the transition score X_i, the school the student is assigned to, and the student's score on the "baccalaureate exam" taken at the end of high school which is the outcome variable Y_i. The quality of school j (treatment dose d_j) is measured by the average transition score of the students attending that school j. The cutoff for admission c_j into a school j is equal to the minimum transition score among the students that are assigned to that school j. The student's preferences on high schools are not observed in the data which makes the RDD fuzzy. For example, a student may have a score greater than the cutoff for the best school in her town but still be assigned to a different school because of her personal preferences. For a transition score X_i, the treatment dose of eligibility $D(X_i)$ is equal to the largest d among those schools with admission cutoff c less than X_i. The treatment dose received D_i coincides with the treatment dose of eligibility $D(X_i)$ for 37% of the students in the sample. Thus, the assignment is quite fuzzy indicating the need for the methods of Section 4 for causal inference beyond ITT effects.

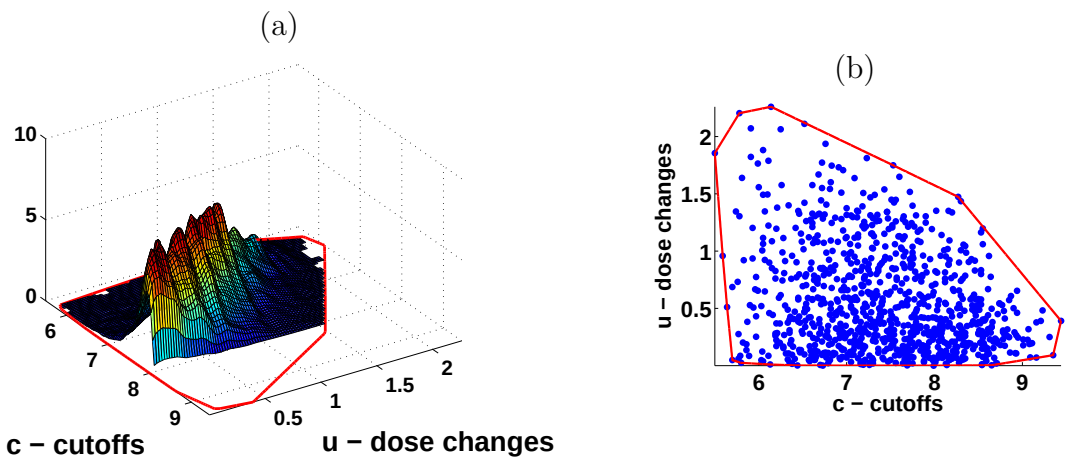
The asymptotic distributions derived in this paper assume independence of first step estimates across cutoffs. Independence across cutoffs is mimicked in the finite sample by matching each individual observation to one single cutoff. After dropping the missing values and those cutoffs with less than 2 observations around it, 939 cutoffs remain with a total of 237,062 observations from 826 schools in 131 towns and 3 years. Only for presentational ease, it is convenient to reduce the dimension of the $\beta(c, d, d')$ function to illustrate counterfactual distributions and non-linearities of treatment effects in three-dimensional graphs. The following restriction is imposed:

$$\beta(c, d, d') \equiv \beta(c, \underbrace{d' - d}_{u}) = \beta(c, u) \tag{20}$$

The ATE estimator proposed in this paper is illustrated at the following example of policy question. Suppose a new charter school is constructed in one of the towns in Romania. The charter school is designed to help students with low test scores and that are currently attending a high school of low quality. More specifically, suppose the charter school admits students by lottery drawing from the national distribution of students with transition score below 8 and that are currently attending a high school of average peer performance below 8. The new charter school has more autonomy and better management than traditional public schools, and its quality is assumed to be equivalent to a high school of average peer performance equal to 8. A student admitted to the charter school that has score X_i and is currently at a school of quality D_i undergoes a change in treatment dose of $U_i = 8 - D_i$. The probability density function of (X_i, U_i) associated with this counterfactual scenario is denoted $\omega(c, u)$ and computed using standard non-parametric kernel density methods (Figure 2(a)). The support of such distribution is more general than the set of finite cutoff-dose values observed in the data (Figure 2(b)). The average effect of the charter school on the academic performance of students is the ATE parameter μ obtained by integrating $\omega(c, u)\beta(c, u)$ over $\mathcal{C}$.

Estimates are computed as following. In the first step estimation, changes in conditional means $\widehat{B}_j$ are obtained by local linear regressions (LLR, $\rho_1 = 1$) with the optimal IK bandwidth of Imbens and Kalyanaraman (2012) and the edge kernel at each cutoff. Out of the 939 cutoffs, there are 147 cutoffs without enough observations to carry out the LLR or the IK bandwidth estimation. For those cutoffs, first step estimates are obtained by a simple Nadaraya-Watson ($\rho_1 = 0$) estimator using all observations linked to each cutoff. In the second step estimation, a bivariate local cubic regression ($\rho_2 = 3$) produces both the estimated function $\widehat{\beta}(c, u)$ and the corrected weights Δ_j. The second step bandwidth is $h_2 = 2.45$ and is chosen to minimize the mean squared error (MSE) of the average treatment effect estimator, $\widehat{\mu} = \sum_j \Delta_j \widehat{B}_j$. The new charter school has an estimated positive average effect that is statistically significant at 1%. Students admitted to the new charter school have their average baccalaureate grade increased by .0868 of a grade point (Table 1).

Figure 2: Charter School Distribution and Support $\mathcal{C}$



Notes: (a) Three dimensional plot of the weighting density $\omega(c, u)$ associated with the charter school example. The contour line on the x,y-plane represents the boundary of set $\mathcal{C}$; (b) Scatter plot of cutoff and dose-change values $\mathcal{C}_K = \{(x_j, u_j)\}_{j=1}^K$ for the 939 cutoffs in the Romanian data. The convex hull enveloping $\mathcal{C}_K$ represents the boundary of the support set $\mathcal{C}$.

Table 1: Average Treatment Effect of Charter School

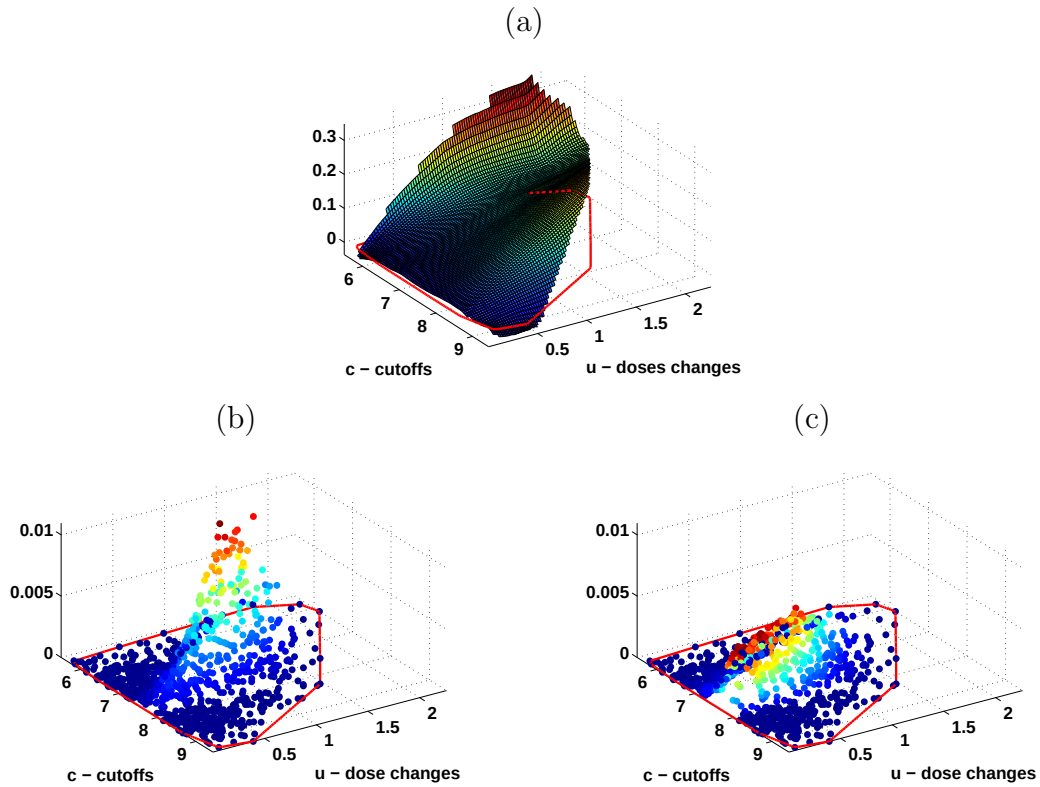
Parameter	Estimate	S.E.
μ	0.0868	0.0239***
μ'	0.0310	0.0177*
$\mu - \mu'$	0.0558	0.0157***

Notes: The parameter μ is the ATE of gaining admission to the charter school on the academic performance of admitted students, and μ is equal to the integral of the treatment effect function $\beta(c, u)$ over the support $\mathcal{C}$ weighted by counterfactual density $\omega(c, u)$. The parameter μ is estimated using the procedures of Section 3. The parameter μ' is an approximation to the true ATE μ, and μ' is equal to the average of $\beta(c_j, u_j)$ weighted by $\omega(c_j, u_j)/\sum_l \omega(c_l, u_l)$, $j = 1, ..., K$. Estimation of μ' uses existing one-cutoff RDD techniques described in Section 2. The difference $\mu - \mu'$ illustrates the error of approximating μ with μ'. All estimates are bias-corrected.

A natural question that arises is whether one could estimate the true ATE μ by simply averaging $\widehat{B}_j$ weighted by the charter school density without the correction scheme introduced in this paper. Such a procedure is consistent for $\mu' = \sum_j \omega(c_j, u_j)\beta(c_j, u_j)/\sum_l \omega(c_l, u_l)$. The estimate $\widehat{\mu'}$ is less than half of the size of $\widehat{\mu}$. The difference $\widehat{\mu} - \widehat{\mu'}$ is statistically different than zero at 1% (Table 1). Simply averaging local effects at existing cutoffs does not give the correct effect of the new charter school policy. Correction weights play a crucial role as they account for the

non-linearities in the treatment effect function $\beta(c, u)$ that are not captured in the simple average of local effects μ'. Figure 3 plots the estimated $\beta(c, u)$ over set $\mathcal{C}$, the corrected weighting scheme used in $\widehat{\mu}$, and the uncorrected weighting scheme used in $\widehat{\mu'}$. The estimator $\widehat{\mu}$ places more weight on the non-linear parts of $\beta(c, u)$ compared to the simple average of local effects $\widehat{\mu'}$. Heterogeneity of treatment effects is a priori unknown and could make the difference $\mu - \mu'$ arbitrarily large. The ATE estimator proposed in this paper performs better than the average of local effects because it accounts for the fact that $\beta(c, u)$ may be any smooth function.

Figure 3: Treatment Effect Function and Weighting Schemes



Notes: (a) Estimated treatment effect function $\widehat{\beta}(c, u)$ over set $\mathcal{C}$; (b) Scatter plot of corrected weighting scheme used to compute the ATE $\widehat{\mu}$ of the charter school policy example; (c) Scatter plot of the uncorrected weighting scheme used to compute $\widehat{\mu'}$ as an approximation to μ.

Theorem 3 says that the class of treatment effect functions must be finite dimensional to obtain identification beyond ITT effects in the fuzzy data of PEU. Following

Assumption 7, the treatment effect function on ever-compliers $\beta_{ec}(c,u)$ is modeled as linear in u – to be consistent with the restriction in Eq. 20 – and quadratic in x to allow for varying marginal effects of ability on returns to school quality.

$$\beta_{ec}(c,u) = \theta_1^{ec}u + \theta_2^{ec}cu + \theta_3^{ec}c^2u$$

Estimates of $\boldsymbol{\theta}^{ec}$ and the ATE of the charter school on ever-compliers μ_{ec} are obtained following the procedure described in Section 4. In the first step, estimates $\widehat{B}_j$ and $\widehat{\widetilde{W}}_j$ are obtained by LLR with the IK bandwidth and edge kernel as before. The quantity $\widehat{\widetilde{W}}_j$ is an estimate for the discontinuity in the average of the vector $[D_i,\ X_iD_i,\ X_i^2D_i]$ conditional on $X_i = c_j$ for every $j = 1,\ldots,K$. In the second step, a first estimate $\widetilde{\boldsymbol{\theta}}^{ec}$ is obtained by regressing $\widehat{B}_j$ on $\widehat{\widetilde{W}}_j$. In the third step, the estimated bias and variance of $\widetilde{\boldsymbol{\theta}}^{ec}$ yield the MSE-optimal weighting scheme $\widetilde{\Omega}$ of Theorem 4. A second estimate $\widehat{\boldsymbol{\theta}}^{ec}$ is obtained by regressing $\widehat{B}_j$ on $\widehat{\widetilde{W}}_j$ using the optimal weights in $\widetilde{\Omega}$. Table 2 reports both of these estimates.

Table 2: Treatment Effects on Ever-compliers

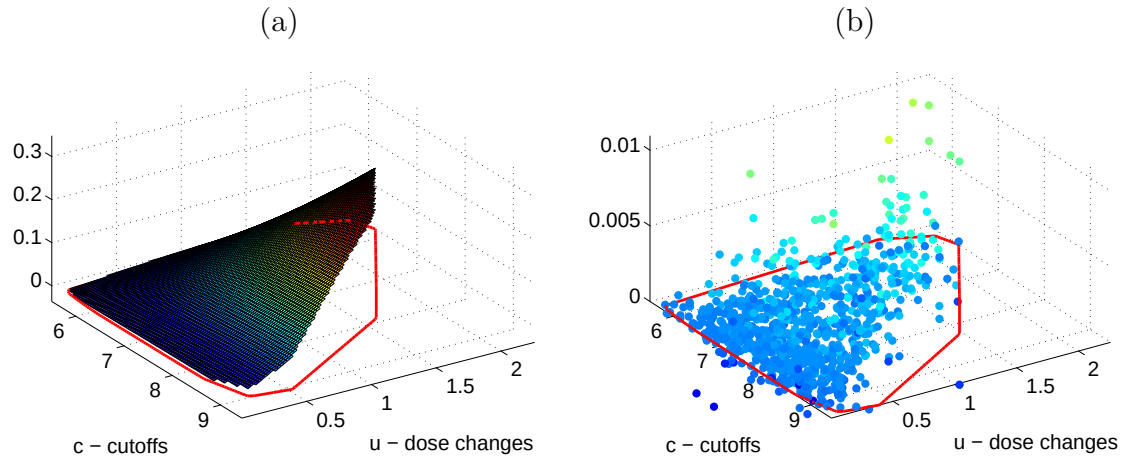
	Equal Weighting		Optimal Weighting	
Parameter	Estimate	S.E.	Estimate	S.E.
θ_1^{ec}	-0.6361	0.3080**	-0.8609	0.1614***
θ_2^{ec}	0.1725	0.0844**	0.2226	0.0472***
θ_3^{ec}	-0.0085	0.0059	-0.0119	0.0035***
μ_{ec}	0.1176	0.0210***	0.0830	0.0119***

Notes: Parametric estimates are obtained according to the procedure described in Section 4 for two choices of weighting matrix Ω: (i) 'Equal Weighting' means $\Omega = I_{K\times K}$ (identity matrix); (ii) 'Optimal Weighting' means that Ω is equal to the inverse of the MSE of $\widetilde{\boldsymbol{\theta}}^{ec}$. The average μ_{ec} is the integral of the estimated parametric $\beta_{ec}(c,u)$ weighted by the charter school weighting density $\omega(c,u)$. All estimates are bias-corrected.

The precision of the parameter estimates is greatly improved by the optimal weighting $\widetilde{\Omega}$, and all parameter estimates become statistically significant at 1% level. The estimated treatment effect function on ever-compliers is plotted in Figure 4(a), and it exhibits a different heterogeneity pattern than ITT effects in Figure 3(a). In both cases, returns to better schooling are increasing in the change of school quality u; however, the return of attending a better school on ever-compliers varies differently with the transition score x than the return of only receiving access to a better school. The ATE $\widehat{\mu}_{ec}$ over ever-compliers that actually attend the charter school is approximately the same the ITT ATE $\widehat{\mu}$. (Tables 1 and 2). Following Equation 18,

the estimator $\widehat{\mu}_{ec}$ is equal to a weighted sum of first step estimates $\widehat{B}_j$. The weighting scheme places more weight on the more precisely estimated $\widehat{B}_j$ at the same time that it accounts for the fuzziness in each cutoff (compare Figure 4(b) to Figure 3(b)).

Figure 4: Treatment Effect Function on Ever-compliers



Notes: (a) Estimated treatment effect function on ever-compliers $\widehat{\beta}_{ec}(c,u) = \widehat{\theta}_1^{ec}u + \widehat{\theta}_2^{ec}cu + \widehat{\theta}_3^{ec}c^2u$. The estimate $\widehat{\boldsymbol{\theta}}^{ec}$ is obtained in three steps according to the procedure described in Section 4 using MSE-optimal weights; (b) Scatter plot of the weights placed on each cutoff in the estimate of $\widehat{\mu}_{ec}$.

6 Conclusion

The difficulty in gathering experimental data in many fields within the social sciences make quasi-experimental techniques such Regression Discontinuity Designs (RDD) extremely important to evaluate policies and social programs using observational data. RDD has been used in a wide range of applications in Economics since the late 1990s. More recently, there has been an increasing number of applications with one forcing variable and multiple cutoffs assigning individuals to heterogeneous treatments. The demand for multi-cutoff RDD methods is constantly growing as the availability of richer data sets continues to increase.

This paper states conditions under which multiple RDD effects are combined to infer average treatment effects (ATE) over the entire range of cutoff values. The proposed estimator is consistent and asymptotically normal for ATEs over the entire support of variation in cutoffs and treatment doses. Asymptotic results are derived

under a large number of observations and cutoffs in the sharp case of non-parametric treatment effect functions. Sufficient conditions on the rate of growth of the number of cutoffs relative to the number of observations are given, and these rate conditions determine the feasible choice set of tuning parameters. This paper also shows that identification in fuzzy RDD with multiple cutoffs is impossible unless the treatment effect function is finite dimensional. It proposes an ATE estimator that is consistent, asymptotically normal, and MSE-optimal under parametric treatment effect functions in the fuzzy case.

The relevance of the ATE estimators proposed in this paper is illustrated with the data of PEU on high school assignment in Romania. Interest lies on the effect of high school quality on academic performance of students. Simple averages of local effects that rely on existing RDD methods fail to estimate the ATE of the counterfactual policy example of a new charter school. There is substantial evidence of non-linearities in the returns to better schooling with respect to the ability level of students. These non-linearities severely bias the simple average of local effects that does not use the correction weighting scheme of the ATE estimator proposed in this paper. Applying the fuzzy RDD methods developed in this paper to the Romanian data reveals causal effects on ever-compliers of similar magnitude but different heterogeneity pattern than Intent-to-Treat effects estimated in the sharp case.

A Appendix

Throughout this appendix, M is used as a generic finite and positive constant in the proofs. For a $p \times q$ matrix A, the norm of A is induced by the Euclidean norm $\|\cdot\|$, i.e. $\|A\| = \max_{x \in \mathbb{R}^q, x \neq 0} \frac{\|Ax\|}{\|x\|}$. Such a matrix norm has the following properties: (i) for a matrix A and a vector x, $\|Ax\| \leq \|A\|\|x\|$; (ii) for matrices A and B such that AB is defined, $\|AB\| \leq \|A\|\|B\|$; (iii) for A invertible, $\|A\|^{-1} \leq \|A^{-1}\|$; and (iv) convergence in the matrix norm is equivalent to convergence of all elements of the matrix. The determinant of matrix A is denoted $\det(A)$.

A.1 Proof of Theorem 1

This theorem is a straightforward generalization of Porter (2003)'s Theorem 3(a), a Central Limit Theorem for the LPR estimator of one discontinuity in a condi-

tional mean function. Such a CLT is restated in Lemma 3 in the Supplemental Appendix B. Note that $\hat{B}_j$ is independent of $\hat{B}_l$ for large n. Applying that CLT gives $\sqrt{nh}\left(\hat{B}_j - B_j\right) \xrightarrow{d} N\left(\mathcal{B}_j, \mathcal{V}_j\right)$ for $\forall j$ under the definitions and conditions of Theorem 1. Among the sufficient conditions used by Lemma 3, there are rate conditions, conditions on the distribution of X_i and on the kernel density that are simply restated in the conditions of Theorem 1. The other two sufficient conditions of Lemma 3 are: (a) $m(x)$ has continuous derivatives wrt x of order $\rho_1 + 1$ in a compact interval centered at c_j but excluding c_j, and side limits at c_j; and (b) continuity of $\zeta^2(x)$ wrt x in a compact interval centered at c_j but excluding c_j, and side limits at c_j. Both of these are implied by Assumption 3. It gives $\nabla_x^{\rho_1+1} m(c_j^+) = \nabla_x^{\rho_1+1} R(c_j, d_j)$, $\nabla_x^{\rho_1+1} m(c_j^-) = \nabla_x^{\rho_1+1} R(c_j, d_{j-1})$, $\zeta^2(c_j^+) = \sigma^2(c_j, d_j)$, $\zeta^2(c_j^-) = \sigma^2(c_j, d_{j-1})$.

□

A.2 Proof of Lemma 2

Define $\overline{\mathcal{C}} = [\underline{\mathcal{X}}, \overline{\mathcal{X}}] \times [\underline{\mathcal{D}}, \overline{\mathcal{D}}] \times [\underline{\mathcal{D}}, \overline{\mathcal{D}}]$. Consider the partition of $\overline{\mathcal{C}}$ made of the set of non-intersecting cubicles $T_n = \{C_1, \ldots, C_M\}$ with $M = n^3$, $n = \{1, 2, \ldots\}$. Each C_j is a half-open cubicle of the form $[x_{l-1}, x_l) \times [y_{m-1}, y_m) \times [z_{o-1}, z_o)$ with sides equal to $(\overline{\mathcal{X}} - \underline{\mathcal{X}})/n$, $(\overline{\mathcal{D}} - \underline{\mathcal{D}})/n$, and $(\overline{\mathcal{D}} - \underline{\mathcal{D}})/n$. Define the sub-collection $U_n = \{C \in T_n : C \subset \mathcal{C}\} = \{A_1, \ldots, A_Q\}$. Since $\mathcal{C}^*$ is dense in $\mathcal{C}$, for every $A_j \in U_n$, find a point $\mathbf{c}_j \in \mathcal{C}^* \cap A_j$ for which $\beta(\mathbf{c}_j)$ is known. The sum $\mu_n = \sum_{j=1}^{Q} \beta(\mathbf{c}_j)\left(\overline{\mathcal{X}} - \underline{\mathcal{X}}\right)\left(\overline{\mathcal{D}} - \underline{\mathcal{D}}\right)^2 / n^3$ converges to $\mu(F^{cont})$ as $n \to \infty$ because $\omega(\mathbf{c})\beta(\mathbf{c})$ is Riemann integrable on $\mathcal{C}$.

□

A.3 Proof of Theorem 2

Let $\widehat{\mu}$ denote $\widehat{\mu}(F^{cont})$ and $\mu = \int_{\mathcal{C}} \omega(\mathbf{c})\beta(\mathbf{c})d(\mathbf{c})$. First, rewrite the normalized and recentered estimator (Lemma 3 in the Supplemental Appendix B shows the details) as

$$\sqrt{Knh_1}\left(\widehat{\mu} - \mu\right) = \sqrt{Knh_1} \sum_j \Delta_j \left(\hat{B}_j - B_j\right) + \sqrt{Knh_1}\left(\sum_j \Delta_j B_j - \mu\right)$$

$$= \sqrt{Knh_1} \sum_j \Delta_j \Bigg\{$$

$$e_1' G_n^{j+} \left[\frac{1}{nh_1} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} \widetilde{H}_i^j \left(Y_i^{j+} - \mathbb{E}[Y_i^{j+} | X_i]\right)\right] \tag{21}$$

$$+e_1' G_n^{j+} \left[\frac{1}{nh_1} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} \widetilde{H}_i^j \mathbb{E}[Y_i^{j+}|X_i] \right.$$

$$\left. - \mathbb{E}\left[\frac{1}{nh_1} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} \widetilde{H}_i^j Y_i^{j+} \right] \quad \right] \tag{22}$$

$$+e_1' G_n^{j+} \mathbb{E}\left[\frac{1}{nh_1} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} \widetilde{H}_i^j Y_i^{j+} \right] \tag{23}$$

$$-e_1' G_n^{j-} \left[\frac{1}{nh_1} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j-} \widetilde{H}_i^j \left(Y_i^{j-} - \mathbb{E}[Y_i^{j-}|X_i]\right) \right] \tag{24}$$

$$-e_1' G_n^{j-} \left[\frac{1}{nh_1} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j-} \widetilde{H}_i^j \mathbb{E}[Y_i^{j-}|X_i] \right.$$

$$\left. - \mathbb{E}\left[\frac{1}{nh_1} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j-} \widetilde{H}_i^j Y_i^{j-} \right] \right] \tag{25}$$

$$\left. -e_1' G_n^{j-} \mathbb{E}\left[\frac{1}{nh_1} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j-} \widetilde{H}_i^j Y_i^{j-} \right] \quad \right\} \tag{26}$$

$$+\sqrt{Knh_1} \left(\sum_j \Delta_j B_j - \mu \right) \tag{27}$$

where the following definitions are used

$H_i^j = [1 \ (X_i - c_j) \ \ldots \ (X_i - c_j)^{\rho_1}]'$, $\widetilde{H}_i^j = \left[1 \ \left(\frac{X_i - c_j}{h_1}\right) \ \ldots \ \left(\frac{X_i - c_j}{h_1}\right)^{\rho_1}\right]'$

$G_n^{j\pm} = \left[\frac{1}{nh_1} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j\pm} \widetilde{H}_i^j \widetilde{H}_i^{j'}\right]^{-1}$, $G^{j\pm} = f(c_j)^{-1} \Gamma_{\pm}^{-1}$

$\Gamma_+ \equiv \Gamma, \quad \Gamma_- \equiv \{(-1)^{j+l} \Gamma_{j,l}\}_{j,l}$

$m(x) = \mathbb{E}[Y_i|X_i = x] - \sum_{j=1}^{K} \mathbb{I}\{c_j \leq x\} B_j$, $\varepsilon_i = Y_i - \mathbb{E}[Y_i|X_i]$

$\varphi^{j+} = \left[m(c_j) + \sum_{l=1}^{j} B_l \quad \nabla_x m(c_j^+) \quad \ldots \quad \nabla_x^{(\rho_1)} m(c_j^+)/\rho_1!\right]'$

$\varphi^{j-} = \left[m(c_j) + \sum_{l=1}^{j-1} B_l \quad \nabla_x m(c_j^-) \quad \ldots \quad \nabla_x^{(\rho_1)} m(c_j^-)/\rho_1!\right]'$

$Y_i^{j+} \equiv Y_i - H_i^{j'} \varphi^{j+} = m(X_i) + \sum_{l=1}^{j} B_l + \varepsilon_i - H_i^{j'} \varphi^{j+}$

$Y_i^{j-} \equiv Y_i - H_i^{j'} \varphi^{j-} = m(X_i) + \sum_{l=1}^{j-1} B_l + \varepsilon_i - H_i^{j'} \varphi^{j-}$

and the definitions of B_j and Γ are given in Equations (3) and (9) in the main text. The idea of this proof is to show that parts (21) and (24) are asymptotically normal; parts (23) and (26) converge in probability to the asymptotic bias; parts (22) and (25) converge in probability to zero; and the integration error (27) converge to zero.

Parts (21) and (24)

$$(21) = \frac{\sqrt{K}}{\sqrt{nh_1}} \sum_j \Delta_j \left[\sum_{i=1}^n k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} e_1' \mathbb{E}\left[G_n^{j+}\right] \widetilde{H}_i^j \varepsilon_i \right] \tag{28}$$

$$+\sqrt{Knh_1} \sum_j \Delta_j \left\{ e_1' \left(G_n^{j+} - \mathbb{E}\left[G_n^{j+}\right]\right) \left[\frac{1}{nh_1} \sum_{i=1}^n k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} \widetilde{H}_i^j \varepsilon_i \right] \right\} \tag{29}$$

Part (28) is a sum of independent zero mean variables for small bandwidth h_1 such that $c_{j-1} + h_1 < c_j - h_1 < c_j + h_1 < c_{j+1} - h_1$ for $\forall j$ (rate condition $Kh_1 = O(1)$ and Assumption 4), and the Lindeberg-Feller CLT is applied. The limiting variance is

$$\begin{aligned}
&\mathbb{V}(28) = \tfrac{K}{nh_1} \textstyle\sum_j \Delta_j^2 n \mathbb{V}\left[k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} e_1' \mathbb{E}\left[G_n^{j+}\right] \widetilde{H}_i^j \varepsilon_i \right] \\
&= K \textstyle\sum_j \Delta_j^2 \frac{1}{h_1} \mathbb{E}\left[k\left(\frac{X_i - c_j}{h_1}\right)^2 \left(e_1' \mathbb{E}\left[G_n^{j+}\right] \widetilde{H}_i^j\right)^2 v_i^{j+} \sigma^2(X_i, d_j) \right] \\
&= K \textstyle\sum_j \Delta_j^2 e_1' \mathbb{E}\left[G_n^{j+}\right] \frac{1}{h_1} \mathbb{E}\left[k\left(\frac{X_i - c_j}{h_1}\right)^2 \widetilde{H}_i^j \widetilde{H}_i^{j\prime} v_i^{j+} \sigma^2(X_i, d_j) \right] \mathbb{E}\left[G_n^{j+}\right]' e_1 \\
&= K \textstyle\sum_j \Delta_j^2 \sigma^2(c_j, d_j) f(c_j) e_1' \mathbb{E}\left[G_n^{j+}\right] \Delta_+ \mathbb{E}\left[G_n^{j+}\right] e_1 + O(h_1) \\
&K \textstyle\sum_j \frac{\Delta_j^2 \sigma^2(c_j, d_j)}{f(c_j)} e_1' \Gamma_+^{-1} \Delta_+ \Gamma_+^{-1} e_1 + O(h_1) \\
&\to \lim_{K,n\to\infty} K \textstyle\sum_j \frac{\Delta_j^2 \zeta^2(c_j^+)}{f(c_j)} e_1' \Gamma_+^{-1} \Delta_+ \Gamma_+^{-1} e_1
\end{aligned}$$

where the $O(h_1)$ term used the facts (i) boundedness of the kernel (Assumption 1); (ii) boundedness of derivatives of f and σ^2 (Assumptions 2 and 5); (iii) f being bounded away from zero (Assumption 2); (iv) $K \sum_j \Delta_j^2 = O(1)$ (Lemma 11 in the Supplemental Appendix B); (vi) $\zeta^2(c_j^+) = \sigma^2(c_j, d_j)$ and $\zeta^2(c_j^-) = \sigma^2(c_j, d_{j-1})$; (vii) $\max_j \|G^{j+} - \mathbb{E}\left[G_n^{j+}\right]\| = O(h_1)$ (Lemma 9 in the Supplemental Appendix B). The limit is equal to

$$\int_{\mathcal{C}} \omega(\mathbf{c})^2 \sigma^2(c, d')/f(c)\, d\mathbf{c} \;\; \left(e_1' \Gamma_+^{-1} \Delta_+ \Gamma_+^{-1} e_1\right) \left(e_1' \mathbf{Q} \Theta\right) \tag{30}$$

where it is used that $\max_j |K\Delta_j/\omega(\mathbf{c}_j) - e_1'\mathbf{Q}\Theta| = o(1)$; the numerical integration weights Δ_j; $\mathbf{Q}$ of Assumption 4; and $\Theta = \lim_{K,n\to\infty} \int_{\mathcal{C}} \Omega_j(\mathbf{c}; h_2) E_j(\mathbf{c}/h_2)\, d\mathbf{c}$; see Part (27) below and Lemma 11 in the Supplemental Appendix B.

To check the Lindeberg-Feller condition, fix $\zeta > 0$ and note that

$$\begin{aligned}
&\textstyle\sum_{i=1}^n \mathbb{E}\Bigg[\left| \frac{\sqrt{K}}{\sqrt{nh_1}} \sum_j \Delta_j k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} e_1' \mathbb{E}\left[G_n^{j+}\right] \widetilde{H}_i^j \varepsilon_i \right|^2 \\
&\qquad\qquad \mathbb{I}\left\{ \left| \frac{\sqrt{K}}{\sqrt{nh_1}} \sum_j \Delta_j k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} e_1' \mathbb{E}\left[G_n^{j+}\right] \widetilde{H}_i^j \varepsilon_i \right| > \zeta \right\} \Bigg] \\
&\leq M \mathbb{E}\left[\left| \frac{K}{\sqrt{Kh_1}} \sum_j \Delta_j v_i^{j+} \right|^2 \mathbb{I}\left\{ M' \left| \frac{K}{\sqrt{Kh_1}} \sum_j \Delta_j v_i^{j+} \right| > \zeta\sqrt{n} \right\} \right]
\end{aligned}$$

$$\leq M\left|\tfrac{K}{\sqrt{Kh_1}}\max_j \Delta_j\right|^2 \mathbb{I}\left\{M'\left|K\max_j \Delta_j\right| > \zeta\sqrt{Knh_1}\right\} \to 0$$

where it is used boundedness of the kernel, ε_i, $\mathbb{E}\left[G_n^{j+}\right]$, $\widetilde{H}_i^j$ within each $[c_j, c_j + h_1)$ and across j; $K\max_j \Delta_j = O(1)$ (Lemma 11 in the Supplemental Appendix B); and the rate condition $Kh_1 = O(1)$. Therefore, part (28) converges in distribution to a normal random variable with limiting variance equal to (30).

Part (29) converges to zero in probability. In fact,

$$|(29)| \leq \sqrt{K}\sum_j |\Delta_j| \ \max_j \left\|G_n^{j+} - \mathbb{E}\left[G_n^{j+}\right]\right\| \ \max_j \left\|\tfrac{1}{\sqrt{nh_1}}\sum_{i=1}^n k\left(\tfrac{X_i - c_j}{h_1}\right) v_i^{j+}\widetilde{H}_i^j \varepsilon_i\right\|$$

Results in the Supplemental Appendix B give a rate for uniform bounds in probability for the second and third terms above. Lemma 9 shows that the second term is $O_P\left(\sqrt{\frac{\log n}{nh_1}}\right)$ and the third term is $O_P(\sqrt{\log n})$. Also, corrected weights are $O(K^{-1})$ uniformly in j which makes the first term $O(\sqrt{K})$. Altogether, $O_P\left(\sqrt{K}\frac{\log n}{\sqrt{nh_1}}\right)$, which is $o_P(1)$ by the rate condition $\frac{\sqrt{K}\log n}{\sqrt{nh_1}} \to 0$. By symmetry of the kernel, a similar reasoning applies to part (24) which is independent of part (21). Therefore, using that $e_1'\Gamma_-^{-1}\Delta_-\Gamma_-^{-1}e_1 = e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1$ and (30) leads to:

$$(21) + (24) \xrightarrow{d} N\left(0; \lim_{K,n\to\infty} K\sum_{j=1}^K \tfrac{\Delta_j^2\left(\zeta^2(c_j^+) + \zeta^2(c_j^-)\right)}{f(c_j)} e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1\right)$$

$$= N\left(0; \left(e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1\right)\left(e_1'\mathbf{Q}\Theta\right)\int_{\mathcal{C}} \omega(\mathbf{c})^2 \tfrac{\sigma^2(c,d') + \sigma^2(c,d)}{f(c)}\, d\mathbf{c}\right)$$

Parts (22) and (25)

$$(22) = \sqrt{Knh_1}\sum_j \Delta_j e_1'\mathbb{E}\left[G_n^{j+}\right]$$

$$\frac{1}{nh_1}\sum_{i=1}^n \left\{k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+}\widetilde{H}_i^j \mathbb{E}[Y_i^{j+}|X_i] - \mathbb{E}\left[k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+}\widetilde{H}_i^j Y_i^{j+}\right]\right\} \tag{31}$$

$$+\sqrt{Knh_1}\sum_j \Delta_j e_1'\left(G_n^{j+} - \mathbb{E}\left[G_n^{j+}\right]\right)$$

$$\frac{1}{nh_1}\sum_{i=1}^n \left\{k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+}\widetilde{H}_i^j \mathbb{E}[Y_i^{j+}|X_i] - \mathbb{E}\left[k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+}\widetilde{H}_i^j Y_i^{j+}\right]\right\} \tag{32}$$

Part (31) is $o_P(1)$ because it has mean zero and a zero limiting variance:

$$\mathbb{V}[(31)] \leq K\sum_j \Delta_j^2 \tfrac{1}{h_1}\mathbb{E}\left[k\left(\tfrac{X_i - c_j}{h_1}\right)^2 \left(e_1'\mathbb{E}\left[G_n^{j+}\right]\widetilde{H}_i^j\right)^2 v_i^{j+}\mathbb{E}[Y_i^{j+}|X_i]^2\right]$$

$$= K\sum_j \Delta_j^2 \tfrac{1}{h_1}\mathbb{E}\left[k\left(\tfrac{X_i - c_j}{h_1}\right)^2 \left(e_1'\mathbb{E}\left[G_n^{j+}\right]\widetilde{H}_i^j\right)^2 v_i^{j+}\left[\tfrac{\nabla_x^{(\rho_1+1)}R(c_j^*, d_j)}{(\rho_1+1)!}(X_i - c_j)^{\rho_1+1}\right]^2\right]$$

$$\leq MK\sum_j \Delta_j^2 \tfrac{1}{h_1}\mathbb{E}\left[k\left(\tfrac{X_i - c_j}{h_1}\right)^2 v_i^{j+}\left(e_1'\mathbb{E}\left[G_n^{j+}\right]\widetilde{H}_i^j\right)^2 h_1^{2(\rho_1+1)}\left|\tfrac{X_i - c_j}{h_1}\right|^{2(\rho_1+1)}\right]$$

$= O(h_1^{2(\rho_1+1)}) = o(1)$

for some constant $c_j^* \in (c_j, c_j + h_1)$ that depends on X_i; and using boundedness of the partial derivative of $R(x,d)$, the density f, the kernel, $\mathbb{E}\left[G_n^{j+}\right]$, and the fact that $K\sum_j \Delta_j^2 = O(1)$ (Assumptions 3 and 5).

Part (32) is also $o_P(1)$

$$(32) \leq \sqrt{Knh_1}\sum_j |\Delta_j| \ \max_j \|G_n^{j+} - \mathbb{E}\left[G_n^{j+}\right]\|$$
$$\max_j \left\| \frac{1}{nh_1}\sum_{i=1}^n \left\{ k\left(\frac{X_i - c_j}{h_1}\right) v_i^{j+} \widetilde{H}_i^j \mathbb{E}[Y_i^{j+}|X_i] - \mathbb{E}\left[k\left(\frac{X_i-c_j}{h_1}\right) v_i^{j+}\widetilde{H}_i^j Y_i^{j+}\right]\right\}\right\|$$

Results in the Supplemental Appendix B give a rate for uniform bounds in probability for the second and third terms above. Lemma 9 shows they are $O_P\left(\sqrt{\frac{\log n}{nh_1}}\right)$. Corrected weights are $O(K^{-1})$ uniformly in j which makes the expression above $O_P\left(\sqrt{K}\frac{\log n}{\sqrt{nh_1}}\right) = o_P(1)$ by the rate condition $\frac{\sqrt{K}\log n}{\sqrt{nh_1}} \to 0$. A similar argument shows that part (25) is $o_P(1)$.

Parts (23) and (26)

$$(23) = \sqrt{Knh_1}\sum_j \Delta_j e_1' \mathbb{E}\left[G_n^{j+}\right]\mathbb{E}\left[\frac{1}{h_1}k\left(\frac{X_i-c_j}{h_1}\right) v_i^{j+}\widetilde{H}_i^j \mathbb{E}[Y_i^{j+}|X_i]\right]$$
$$+\sqrt{Knh_1}\sum_j \Delta_j e_1' \left(G_n^{j+} - \mathbb{E}\left[G_n^{j+}\right]\right)\mathbb{E}\left[\frac{1}{nh_1}\sum_{i=1}^n k\left(\frac{X_i-c_j}{h_1}\right) v_i^{j+}\widetilde{H}_i^j Y_i^{j+}\right]$$
$$= \sqrt{Knh_1}\sum_j \Delta_j e_1' \Bigg\{$$
$$\mathbb{E}\left[G_n^{j+}\right]\mathbb{E}\left[v_i^{j+}\frac{1}{h_1}k\left(\frac{X_i-c_j}{h_1}\right)\widetilde{H}_i^j \frac{\nabla_x^{(\rho_1+1)}R(c_j,d_j)}{(\rho_1+1)!}(X_i-c_j)^{\rho_1+1}\right]$$
$$+\mathbb{E}\left[G_n^{j+}\right]\mathbb{E}\left[v_i^{j+}\frac{1}{h_1}k\left(\frac{X_i-c_j}{h_1}\right)\widetilde{H}_i^j \frac{\nabla_x^{(\rho_1+2)}R(c_j^*,d_j)}{(\rho_1+2)!}(X_i-c_j)^{\rho_1+2}\right]$$
$$+\left(G_n^{j+} - \mathbb{E}\left[G_n^{j+}\right]\right)\mathbb{E}\left[\frac{1}{nh_1}\sum_{i=1}^n k\left(\frac{X_i-c_j}{h_1}\right) v_i^{j+}\widetilde{H}_i^j Y_i^{j+}\right] \Bigg\}$$
$$= \sqrt{Knh_1}\sum_j \Delta_j e_1' \Bigg\{$$
$$h_1^{\rho_1+1}\left[G^{j+}\frac{\nabla_x^{(\rho_1+1)}R(c_j,d_j)}{(\rho_1+1)!}f(c_j)\gamma^* + O\left(h_1\right)\right]$$
$$+O\left(h_1^{\rho_1+2}\right)$$
$$+O_P\left(h_1^{\rho_1+1}\frac{\log n}{nh_1}\right) \Bigg\}$$
$$= C\lim_{K,n\to\infty}\sum_j \Delta_j \frac{\nabla_x^{(\rho_1+1)}R(c_j,d_j)}{(\rho_1+1)!}e_1'\Gamma^{-1}\gamma^* + O\left(\sqrt{Knh_1}h_1^{\rho_1+2}\right)$$

where it is used that (i) the rate condition $\sqrt{Knh_1}h_1^{\rho_1+1} \to C \in (0,\infty)$; the uniform bounds derived in Supplemental Appendix B (Lemma 9) (ii) $\max_j \|G^{j+} - \mathbb{E}\left[G_n^{j+}\right]\| = O(h_1)$; (iii) $\max_j \|G_n^{j+} - \mathbb{E}\left[G_n^{j+}\right]\| = O_P\left(\sqrt{\frac{\log n}{nh_1}}\right)$; (iv) $\sum_j |\Delta_j| = O(1)$;

(v) boundedness of derivatives, density of X and $\mathbb{E}\left[G_n^{j+}\right]$. The limit is equal to $C\int_{\mathcal{C}}\omega(\mathbf{c})\frac{\nabla_x^{(\rho_1+1)}R(c,d')}{(\rho_1+1)!}e_1'\Gamma^{-1}\gamma^*\ d\mathbf{c}$, see Part (27) below and Lemma 11 in the Supplemental Appendix B. A similar idea applies to part (26) which leads to

$$(23)+(26)\xrightarrow{p}C\left(e_1'\Gamma^{-1}\gamma^*\right)\int_{\mathcal{C}}\omega(\mathbf{c})\frac{\nabla_x^{(\rho_1+1)}R(c,d')-(-1)^{\rho_1+1}\nabla_x^{(\rho_1+1)}R(c,d)}{(\rho_1+1)!}\ d\mathbf{c}$$

Part (27)

If the schedule of cutoff-doses is asymptotically dense (Assumption 4), the number of bounded derivatives on the treatment effect function $\beta(c,d,d')$ (Assumption 5) determines the rate at which the integration error converges to zero (Lemma 11, Supplemental Appendix B). This leads to:

$$(27)=\sqrt{Knh_1}\left(\sum_j\Delta_j\beta_j-\int_{\mathcal{C}}\omega(\mathbf{c})\beta(\mathbf{c})d(\mathbf{c})\right)=\sqrt{Knh_1}O(h_2^{\rho_2+1})=o(1)$$

by the rate condition $\sqrt{Knh_1}h_2^{\rho_2+1}=o(1)$.

□

A.4 Proof of Theorem 3

Define $\delta_{j,l}=\mathbb{I}\{\mathcal{U}_i(c_j)=d_l\}$. Assumption 6 (no-ever-defiance) implies the following facts: (i) $\mathbb{P}\left[\delta_{j-1,l}=0,\ \delta_{j,l}=1\right]=0$ for $\forall l\neq j$; (ii) $\mathbb{P}\left[\delta_{j-1,l}=1,\ \delta_{j,l}=0\right]=0$ for $l=j$; (iii) $\mathbb{P}\left[\delta_{j-1,l}=1,\ \delta_{j,l}=0,\ \delta_{j,u}=1\right]=0$ for $\forall u\neq j$ and $u\neq l$.

Fix a small $e>0$ and use fact (i) to obtain:

$\mathbb{E}[Y_i|X_i=c_j+e]=\sum_{l=0}^K\mathbb{E}\left[\delta_{j,l}Y_i(d_l)|X_i=c_j+e\right]$
$=\sum_{l=0}^K\mathbb{E}\left[Y_i(d_l)|X_i=c_j+e,\ \delta_{j,l}=1,\ \delta_{j-1,l}=1\right]\mathbb{P}\left[\delta_{j,l}=1,\ \delta_{j-1,l}=1|X_i=c_j+e\right]$
$+\sum_{l=0}^K\mathbb{E}\left[Y_i(d_l)|X_i=c_j+e,\ \delta_{j,l}=1,\ \delta_{j-1,l}=0\right]\mathbb{P}\left[\delta_{j,l}=1,\ \delta_{j-1,l}=0|X_i=c_j+e\right]$
$=\sum_{l=0}^K\mathbb{E}\left[Y_i(d_l)|X_i=c_j+e,\ \delta_{j,l}=1,\ \delta_{j-1,l}=1\right]\mathbb{P}\left[\delta_{j,l}=1,\ \delta_{j-1,l}=1|X_i=c_j+e\right]$
$+\mathbb{E}\left[Y_i(d_j)|X_i=c_j+e,\ \delta_{j,j}=1,\ \delta_{j-1,j}=0\right]\mathbb{P}\left[\delta_{j,j}=1,\ \delta_{j-1,j}=0|X_i=c_j+e\right]$

Take the limit as $e\downarrow 0$, and use that $\{\delta_{j,l}=1,\ \delta_{j-1,l}=1\}$ and $\{\delta_{j,j}=1,\ \delta_{j-1,j}=0\}$ are finite unions of sets of the form $\{\mathcal{U}_i=\bar{\mathcal{U}}\}$, $\bar{\mathcal{U}}\in\mathcal{U}^*$, that is, measurable subsets of $\mathcal{A}$ for which the conditional expectation and probability are continuous functions of x (Assumption 6).

$\lim_{e\downarrow 0}\mathbb{E}[Y_i|X_i=c_j+e]$
$=\sum_{l=0}^K\mathbb{E}\left[Y_i(d_l)|X_i=c_j,\ \delta_{j,l}=1,\ \delta_{j-1,l}=1\right]\mathbb{P}\left[\delta_{j,l}=1,\ \delta_{j-1,l}=1|X_i=c_j\right]$
$+\mathbb{E}\left[Y_i(d_j)|X_i=c_j,\ \delta_{j,j}=1,\ \delta_{j-1,j}=0\right]\mathbb{P}\left[\delta_{j,j}=1,\ \delta_{j-1,j}=0|X_i=c_j\right]$
$=\sum_{l=0}^K\mathbb{E}\left[Y_i(d_l)|X_i=c_j,\ \delta_{j,l}=1,\ \delta_{j-1,l}=1\right]\mathbb{P}\left[\delta_{j,l}=1,\ \delta_{j-1,l}=1|X_i=c_j\right]$
$+\sum_{l=0,l\neq j}^K\mathbb{E}\left[Y_i(d_j)|X_i=c_j,\ \delta_{j,j}=1,\ \delta_{j-1,l}=1\right]\mathbb{P}\left[\delta_{j,j}=1,\ \delta_{j-1,l}=1|X_i=c_j\right]$

Similarly, use fact (ii) for the left-hand-side limit, $\lim_{e\downarrow 0}\mathbb{E}[Y_i|X_i=c_j-e]$

$= \sum_{l=0}^{K} \mathbb{E}\left[Y_i(d_l)|X_i = c_j,\ \delta_{j,l} = 1,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,l} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right]$
$+ \sum_{l=0,l\neq j}^{K} \mathbb{E}\left[Y_i(d_j)|X_i = c_j,\ \delta_{j,l} = 0,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,l} = 0,\ \delta_{j-1,l} = 1|X_i = c_j\right]$
Use fact (iii) to get
$= \sum_{l=0}^{K} \mathbb{E}\left[Y_i(d_l)|X_i = c_j,\ \delta_{j,l} = 1,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,l} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right]$
$+ \sum_{l=0,l\neq j}^{K} \mathbb{E}\left[Y_i(d_l)|X_i = c_j,\ \delta_{j,j} = 1,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,j} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right]$
The difference between right and left hand side limits is B_j
$= \sum_{l=0,l\neq j}^{K} \mathbb{E}\left[Y_i(d_j) - Y_i(d_l)|X_i = c_j, \delta_{j,j} = 1, \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,j} = 1, \delta_{j-1,l} = 1|X_i = c_j\right]$
$= \sum\limits_{l=0,l\neq j}^{K} \beta_{ec}(c_j, d_l, d_j)\mathbb{P}\left[\delta_{j,j} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right]$
Next, it is shown that $\mathbb{P}\left[\delta_{j,j} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right] = \omega_{j,l}$, for $l \neq j$.
$\mathbb{P}\left[\delta_{j,j} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right] = \mathbb{P}[\delta_{j,l} = 0, \delta_{j-1,l} = 1|X_i = c_j]$
$= \mathbb{P}[\delta_{j,l} = 0|X_i = c_j] - \mathbb{P}[\delta_{j-1,l} = 0|X_i = c_j]$
$= \lim_{e\downarrow 0} \left\{\mathbb{P}[\mathcal{U}_i(c_j) \neq d_l|X_i = c_j + e] - \mathbb{P}[\mathcal{U}_i(c_{j-1}) \neq d_l|X_i = c_j - e]\right\}$
$= \lim_{e\downarrow 0} \left\{\mathbb{P}[D_i = d_l|X_i = c_j - e] - \mathbb{P}[D_i = d_l|X_i = c_j + e]\right\}$ where facts (i) and (ii) are used. This proves the first part of the Lemma.

If β_{ec} belongs to the class of functions of Assumption 7, then $B_j = \widetilde{W}_j\boldsymbol{\theta}_0$. If the matrix $\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}} = \sum\limits_j \widetilde{W}_j\widetilde{W}_j'$ is invertible, then the second part of the Lemma follows.

Conversely, suppose that the $p > K$ elements in $\{\beta_{ec}(c_j, d_l, d_j)$ for $(j, l) : \omega_{j,l} > 0\}$ are identified for every fuzzy assignment $\tilde{\mathbf{c}}_1 = (c_1, d_0, d_1)$, ..., $\tilde{\mathbf{c}}_p = (c_K, d_{K-1}, d_K)$. Identification means that there is an unique solution to the following constrained linear system:

$$\begin{bmatrix} B_1 \\ \vdots \\ B_K \end{bmatrix} = \begin{bmatrix} \omega_{1,0} & \dots & \omega_{1,K} & 0 & 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & \omega_{2,0} & \dots & \omega_{2,K} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & & & & & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & \omega_{K,0} & \dots & \omega_{K,K-1} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\text{such that } (\beta_1, \dots, \beta_p) \in \mathcal{G}$$

The $K \times p$ matrix of coefficients has rank equal to K because the assignment is fuzzy. Since $p > K$, the unconstrained system has infinitely many nonzero solutions of the form $\mathbf{b} = \mathbf{b}^p + \sum_{m=1}^{p-K} \lambda_m \mathbf{b}_m^s$ for any $(\lambda_1, \dots, \lambda_{p-K}) \in \mathbb{R}^{p-K}$, where $\{\mathbf{b}_m^s\}_{m=1}^{p-K}$ are the basis vectors of the null-space of the unconstrained system, and $\mathbf{b}^p$ is a particular solution. The constrained system has one unique solution $\mathbf{b}^* \in \mathcal{G}$, however, $\mathbf{b}^* + \mathbf{b}_m^s \notin \mathcal{G}\ \ \forall m$. This implies that $\mathbf{b}_m^s \notin \mathcal{G}\ \ \forall m$ because $\mathcal{G}$ is a vector subspace of $\mathbb{R}^p$. This is a set of $p - K$ linearly independent vectors in $\mathbb{R}^p$ not in $\mathcal{G}$. Therefore, the $dim\mathcal{G} \leq p - (p - K) = K$, and third part of the Lemma follows. □

A.5 Proof of Theorem 4

This proof uses a Central Limit Theorem for the LPR estimator of the difference in side-limits of a conditional mean function of the vector $\mathbf{Y}_i$ given X_i at $X_i = c_j$. The result is well known in the scalar case (Porter (2003)), and it is generalized to the multivariate case in Lemma 3 in the Supplemental Appendix B. Such result is applied on $\mathbf{Y}_i = [Y_i \quad \boldsymbol{\mathcal{W}}(c_j, D_i)']'$ to show $\sqrt{nh}\left(\hat{\mathbf{J}}_j - \mathbf{J}_j\right) \to N\left(\boldsymbol{\Psi}_j; \boldsymbol{\Phi}_j\right)$ using the conditions and definitions of Theorem 4. Among the sufficient conditions used by Lemma 3, there are rate conditions, conditions on the distribution of X_i and on the kernel density that are simply restated in the conditions of Theorem 4. It remains to show the other two sufficient conditions of Lemma 3 that are: (a) $\mathbf{m}(x)$ has continuous derivatives wrt x of order $\rho_1 + 1$ in a compact interval centered at c_j but excluding c_j, and side limits at c_j; and (b) continuity of $\boldsymbol{\zeta}^2(x)$ wrt x in a compact interval centered at c_j but excluding c_j, and side limits at c_j.

For (a), note that, in the fuzzy case, the mean of Y_i and $\boldsymbol{\mathcal{W}}(c_j, D_i)$ conditional on X_i is a sum of the means of potential outcomes $Y_i(d)$ and $\boldsymbol{\mathcal{W}}(c_j, d)$ for various dosages d conditional on sets of the form $\{\mathcal{U}_i(c_j) = d_l\}$ weighted by conditional probabilities of the same sets (see proof of Theorem 3). Assumption 8 implies that such conditional means and probabilities above are smooth functions of x and side-limits exist at $x = c_j$. Similarly, for (b), the conditional covariance of $(Y_i, \boldsymbol{\mathcal{W}}(c_j, D_i))$ is a function of sums of the first and second moments of potential outcomes $Y(d)$ for various dosages conditional on sets of the form $\{\mathcal{U}_i(c_j) = d_l\}$ weighted by conditional probabilities of the same sets. Assumption 8 ensures continuity of $\boldsymbol{\zeta}^2(x)$ wrt x and existence of side-limits at c_j. Therefore, $\sqrt{nh}\left(\hat{\mathbf{J}}_j - \mathbf{J}_j\right) \to N\left(\boldsymbol{\Psi}_j; \boldsymbol{\Phi}_j\right)$ for every $j = 1, \ldots, K$.

Next, note that

$$\lim_{e \downarrow 0} \left\{ \mathbb{E}\left[\boldsymbol{\mathcal{W}}(c_j, D_i) | X_i = c_j + e\right] - \mathbb{E}\left[\boldsymbol{\mathcal{W}}(c_j, D_i) | X_i = c_j - e\right] \right\}$$

$= \sum_{l=0, l \neq j}^{K} \left\{ \boldsymbol{\mathcal{W}}(c_j, d_j) - \boldsymbol{\mathcal{W}}(c_j, d_l) \right\} \omega_{j,l} = \widetilde{W}_j$ which means that $\mathbf{J}_j = [B_j \quad \widetilde{W}_j']'$. Then, the CLT above implies

$$\sqrt{nh_1}\left(\widehat{B}_j - \widehat{\widetilde{W}}_j' \boldsymbol{\theta}_0^{ec}\right) = [1 \quad -\boldsymbol{\theta}_0^{ec\prime}] \sqrt{nh_1} \begin{pmatrix} \widehat{B}_j - B_j \\ \widehat{\widetilde{W}}_j - \widetilde{W}_j \end{pmatrix} \xrightarrow{d} N\left(C\ \mathcal{B}_j^{ec}; \mathcal{V}_j^{ec}\right) \quad \forall j$$

and

$\widehat{\widetilde{W}}_j \xrightarrow{p} \widetilde{W}_j \ \ \forall j$, or, $\widehat{\widetilde{\mathbf{W}}} \xrightarrow{p} \widetilde{\mathbf{W}}$.

For large n and fixed K, $(\widehat{B}_j, \widehat{\widetilde{W}}_j) \perp (\widehat{B}_l, \widehat{\widetilde{W}}_l)$ for $l \neq j$. Therefore,

$$\sqrt{nh_1}\left(\widehat{\mathbf{B}} - \widehat{\widetilde{\mathbf{W}}}\boldsymbol{\theta}_0^{ec}\right) \xrightarrow{d} N\left(C\ \boldsymbol{\mathcal{B}}^{ec}; \boldsymbol{\mathcal{V}}^{ec}\right)$$

and

$$\sqrt{nh_1}\left(\widehat{\boldsymbol{\theta}}^{ec} - \boldsymbol{\theta}_0^{ec}\right) = \left(\widehat{\widetilde{\mathbf{W}}}'\Omega\widehat{\widetilde{\mathbf{W}}}\right)^{-1}\widehat{\widetilde{\mathbf{W}}}'\Omega\sqrt{nh_1}\left(\widehat{\mathbf{B}} - \widehat{\widetilde{\mathbf{W}}}\boldsymbol{\theta}_0^{ec}\right)$$

$$\xrightarrow{d} N\left(C\ \left(\widetilde{\mathbf{W}}'\Omega\widetilde{\mathbf{W}}\right)^{-1}\widetilde{\mathbf{W}}'\Omega\boldsymbol{\mathcal{B}}^{ec}; \left(\widetilde{\mathbf{W}}'\Omega\widetilde{\mathbf{W}}\right)^{-1}\widetilde{\mathbf{W}}'\Omega\boldsymbol{\mathcal{V}}^{ec}\Omega\widetilde{\mathbf{W}}\left(\widetilde{\mathbf{W}}'\Omega\widetilde{\mathbf{W}}\right)^{-1}\right)$$

and the choice $\Omega = \left(C\ \boldsymbol{\mathcal{B}}^{ec}\boldsymbol{\mathcal{B}}^{ec\prime} + \boldsymbol{\mathcal{V}}^{ec}\right)^{-1}$ minimizes the MSE of $\sqrt{nh_1}\left(\widehat{\boldsymbol{\theta}}^{ec} - \boldsymbol{\theta}_0^{ec}\right)$.

□

References

Agarwal, S., S. Chomsisengphet, N. Mahoney, and J. Stroebel (2016): "Do Banks Pass Through Credit Expansions to Consumers Who Want To Borrow?," Working Paper Available at SSRN: http://ssrn.com/abstract=2617116.

Angrist, J. (2004): "Treatment Effect Heterogeneity in Theory and Practice," *The Economic Journal*, 114, C52–C83.

Angrist, J., and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *The Quarterly Journal of Economics*, 114(2), 533–575.

Angrist, J., 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.

Angrist, J. D., and J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bajari, P., H. Hong, M. Park, and R. Town (2011): "Regression Discontinuity Designs With An Endogenous Forcing Variable And An Application To Contracting In Health Care," *NBER Working Paper 17643*.

Bertanha, M., and G. Imbens (2014): "External Validity in Fuzzy Regression Discontinuity Designs," *NBER Working Paper 20773*.

Black, D. A., J. Galdo, and J. A. Smith (2007): "Evaluating the Worker Profiling and Reemployment Services System Using a Regression Discontinuity Approach," *The American Economic Review*, 97(2), 104–107.

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

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

Cattaneo, M., L. Keele, R. Titiunik, and G. Vazquez-Bare (2016): "Interpreting Regression Discontinuity Designs with Multiple Cutoffs," *Journal of Politics, forthcoming*.

Chen, S., and W. Van der Klaauw (2008): "The Work Disincentive Effects of the Disability Insurance Program in the 1990s," *Journal of Econometrics*, 142(2), 757–784.

De Giorgi, G., A. Drenik, and E. Seira (2015): "Moral Hazard in the Credit Card Market," *Working Paper, Stanford University*.

De La Mata, D. (2012): "The Effect of Medicaid Eligibility on Coverage, Utilization, and Children's Health," *Health Economics*, 21(9), 1061–1079.

Dobkin, C., and F. Ferreira (2010): "Do School Entry Laws Affect Educational Attainment And Labor Market Outcomes?," *Economics of Education Review*, 29(1), 40–54.

Dong, Y. (2016a): "An Alternative Assumption to Identify LATE in Regression Discontinuity Designs," *Working Paper, University of California, Irvine*.

——— (2016b): "Jump or Kink? Identifying Education Effects by Regression Discontinuity Design without the Discontinuity," *Working Paper, University of California, Irvine*.

Dong, Y., and A. Lewbel (2015): "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," *The Review of Economics and Statistics*, 97(5), 1081–1092.

Duflo, E., P. Dupas, and M. Kremer (2011): "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 101(5), 1739–1774.

Egger, P., and M. Koethenbuerger (2010): "Government Spending and Legislative Organization: Quasi-experimental Evidence from Germany," *American Economic Journal: Applied Economics*, 2(4), 200–212.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications:*

Monographs on Statistics and Applied Probability, Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Frandsen, R., M. Frölich, and B. Melly (2012): "Quantile Treatment Effects In The Regression Discontinuity Design," *Journal of Econometrics*, 168(2), 382–395.

Garibaldi, P., F. Giavazzi, A. Ichino, and E. Rettore (2012): "College Cost and Time to Obtain a Degree: Evidence from Tuition Discontinuities," *The Review of Economics and Statistics*, 94(3), 699–711.

Hahn, J., P. Todd, and W. Van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1), 201–209.

Hardle, W., and A. W. Bowman (1988): "Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands," *Journal of the American Statistical Association*, 83(401), 102–110.

Hastings, J. S., C. A. Neilson, and S. D. Zimmerman (2013): "Are Some Degrees Worth More Than Others? Evidence From College Admission Cutoffs In Chile," *NBER Working Paper 19241*.

Hoekstra, M. (2009): "The Effect of Attending the Flagship State University on Earnings: A Discontinuity-based Approach," *The Review of Economics and Statistics*, 91(4), 717–724.

Hoxby, C. (2000): "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *The Quarterly Journal of Economics*, 115(4), 1239–1285.

Imbens, G., and K. Kalyanaraman (2012): "Optimal Bandwidth Choice For The Regression Discontinuity Estimator," *The Review of Economic Studies*, 79(3), 933–959.

Imbens, G. W., and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide To Practice," *Journal of Econometrics*, 142(2), 615–635.

Imbens, G. W., and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *The Review of Economic Studies*, 64(4), 555–574.

Lazear, E. (2001): "Educational Production," *Quarterly Journal of Economics*, 116(3), 777–803.

McCrary, J. (2008): "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142(2), 698–714.

McCrary, J., and H. Royer (2011): "The Effect of Female Education on Fertility and Infant Health: Evidence from School Entry Policies Using Exact Date of Birth," *American Economic Review*, 101(1), 158–195.

Neumann, M. H., and J. Polzehl (1998): "Simultaneous Bootstrap Confidence Bands in Nonparametric Regression," *Journal of Nonparametric Statistics*, 9(4), 307–333.

Otsu, T., K.-L. Xu, and Y. Matsushita (2015): "Empirical Likelihood for Regression Discontinuity Design," *Journal of Econometrics*, 186(1), 94–112.

Pop-Eleches, C., and M. Urquiola (2013): "Going to a Better School: Effects and Behavioral Responses," *American Economic Review*, 103(4), 1289–1324.

Porter, J. (2003): "Estimation in the Regression Discontinuity Model," *Unpublished Manuscript, University of Wisconsin at Madison.*

Rokkanen, M. (2015): "Exam Schools, Ability, and the Effects of Affirmative Action: Latent Factor Extrapolation in the Regression Discontinuity Design," *Unpublished Manuscript, Columbia University.*

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66(5), 688.

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.