Regression Discontinuity Design with Many Thresholds*

Marinho Bertanha†

May 29, 2015

Abstract

In recent years, numerous studies have employed regression discontinuity designs with many cutoffs assigning individuals to heterogeneous treatments. A common practice is to normalize all of the cutoffs to zero and estimate only one effect. This procedure identifies the average of local treatment effects weighted by the observed relative density of individuals at the existing cutoffs. However, researchers often want to make inferences on more meaningful average treatment effects (ATE) computed over general counterfactual distributions of individuals rather than simply the observed distribution of individuals local to existing cutoffs. In this paper, we propose a root-n consistent and asymptotically normal estimator for such ATEs when heterogeneity follows a non-parametric smooth function of cutoff characteristics. In the case of parametric heterogeneity, observations are optimally combined to minimize the mean squared error of the ATE estimator. Inference results are also provided for the fuzzy regression discontinuity case, where the parametric heterogeneity assumption yields identification of treatment effects on individuals who comply with at least one of the multiple treatments.

Most Recent Version Available at:

www.stanford.edu/∼bertanha/Bertanha_JMP.pdf

*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, and UCSD. I thank Arun Chandrasekhar, Michael Dinerstein, Ivan Korolev, Michael Leung, Huiyu Li, Jessie Li, 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.

†Stanford University, Department of Economics, 579 Serra Mall, Stanford-CA 94305. Email: bertanha@stanford.edu. Website: www.stanford.edu/∼bertanha.

1 Introduction

One of the fundamental problems in economic analyses with observational data is that we do not see the counterfactual scenario needed to make causal inferences. When the researcher has a theoretical relationship among variables in mind, it is the task of econometrics to derive minimal conditions for this relationship such that the causal effect is identified and feasible to estimate (White and Chalak (2013), Heckman and Vytlacil (2007)). 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 (e.g. Black, Galdo, and Smith (2007), Egger and Koethenbuerger (2010), De La Mata (2012), Pop-Eleches and Urquiola (2013)). Existing one-cutoff RDD methods can be applied to each cutoff individually but yield local effects that are estimated using only a few observations near each cutoff. Researchers often prefer one takeaway summary effect that can be 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. If treatment effects are heterogeneous, 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 (2015)). Such an average effect would be a meaningful summary measure only in two cases: (i) all local treatment effects are identical and the weighting scheme does not matter; (ii) the researcher is interested in the average effect of heterogeneous treatments only on the observed distribution of individuals near the existing cutoffs of such treatments. 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]

The inference procedure of our paper estimates an average treatment effect (ATE) that is a more valuable summary measure than the average effect identified by the normalization procedure because the weighting scheme can be explicitly chosen according to the counterfactual scenario of interest. For applications with substantial variation in cutoff values, we propose a corrected weighting scheme that allows for inference on ATEs computed over an explicitly chosen distribution of individuals, including individuals in between existing cutoffs.

We motivate our framework for RDD with many thresholds using a simple example based on Pop-Eleches and Urquiola (2013), PEU from now on. They study the assignment of students to more or less elite high schools based on test scores where every town has its own set of minimum admission cutoffs scores and each town's cutoffs vary from year to year. 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 and on the behavior of parents and teachers. The economic logic of their application can be briefly summarized as follows.[2] Suppose there is a central planner who assigns students to high schools based on the students' scores on a placement test. High schools have limited capacity and can be ranked by some measure of quality. 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 or threshold) for admission to each high school j in each town.[3] The quality of high school j is

[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) finds 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.

[2] The purpose of the simple example in the paragraph above is to introduce the reader to the framework of RDD with multiple cutoffs. The details of PEU's application are described in our application section.

[3] For the sake of exposition, we assume for now that students attend the best high school available to them based on their score and the cutoffs that apply to them. In fact, some students may choose to attend a lesser high school so that we have a fuzzy rather than a sharp RDD. We examine the fuzzy case later in the paper.

denoted d_j, and different school qualities expose students to different treatment doses across cutoffs c_j. We are interested in the treatment effect of high school quality on the student i's academic achievement Y_i (outcome variable). This effect is not immediately identified because we never observe the same students attending high schools of different qualities. 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. By comparing students with test scores just below the cutoff to students with scores just above the cutoff, RDD allows identification of the impact of school quality on the average academic achievement of those students with test scores equal to c_j. We denote such local treatment effect by the average $\beta_j = \mathbb{E}[Y_i(d_j) - Y_i(d_{j-1})|X_i = c_j]$, where $Y_i(d_j)$ is the potential academic achievement student i has if attending a high school of quality d_j, and $\mathbb{E}[.]$ denotes expected value over the distribution of students. 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. Nevertheless, there are numerous other examples of RDD with multiple thresholds 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 thresholds. De La Mata (2012) estimates the effects of Medicaid benefits on healthcare utilization, where Medicaid eligibility is triggered by income thresholds that vary across states. De Giorgi, Drenik, and Seira (2015) look at the effect of credit approval on the probability of default, where banks use thresholds on credit scores to make 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, e.g. 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.

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 both in the sharp and fuzzy cases were formalized by Hahn, Todd, and Van der Klaauw (2001) who proposed estimation by local linear regression and derived its asymptotic distribution. Local polynomial regressions are known for low order bias at boundary points and rate optimality (Fan and Gijbels (1996), Porter (2003)). 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 (2014)).

More closely related to our contribution is the study of treatment effect extrapolation, e.g. Angrist (2004), Bertanha and Imbens (2015), Dong and Lewbel (2015), Angrist and Rokkanen (2013), and Rokkanen (2014). 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. Our results differ from theirs because we use the variation of multiple cutoffs and doses to identify ATEs over distributions of individuals in between and at cutoffs without restricting the heterogeneity of treatment effects. 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 to estimate an average effect depends on how comparable the researcher believes these effects are. We consider three cases of heterogeneity

assumptions researchers could make about the comparability of treatment effects. The cases are presented in increasing order of structure. In the first case of heterogeneity assumptions, the researcher does not believe changes in treatment doses $d_{j-1} \rightarrow d_j$ are quantitatively comparable across cutoffs c_j. This would be the case in the high school assignment example, if the quality of different schools could not be credibly summarized in one metric d. Another example is the RDD setup studied in Hastings, Neilson, and Zimmerman (2013), where different score cutoffs assign students to different degree programs in Universities in Chile. One cutoff could switch students from Physics to Engineering and a second cutoff from Engineering to Economics, and it is difficult to summarize these treatment dose changes in one metric.

In the second case of heterogeneity assumptions, the researcher believes that treatment doses are quantitatively comparable across cutoffs, and treatment effects vary smoothly with changes in treatment dose. In the high school assignment example, PEU measures high school quality using average student performance in each school. They find behavioral evidence that parents and teachers perceive the schools' quality based on this measure. Another example of multiple cutoffs with comparable treatment doses is the case where there is one type of treatment triggered by one cutoff, but this one cutoff varies across subpopulations (e.g. towns, years, states). This is the case of Medicaid benefits for children studied in De La Mata (2012), where each state (subpopulation) has its own income threshold for Medicaid coverage eligibility (one treatment).

In the third case of heterogeneity assumptions, the researcher is willing to specify a parametric functional form for the treatment effect function $\beta(x, d, d') = \mathbb{E}[Y_i(d') - Y_i(d) | X_i = x]$. Economic theory or a priori knowledge guides the choice of a functional form that credibly summarizes the heterogeneity of treatment effects. In the high school assignment example, researchers could assume that returns to better education follow a polynomial function of scores and school quality to test for varying marginal returns. In addition to average effects, the researcher could be interested in structural parameters of an underlying model that

predicts a parametric treatment effect function. In a class size application like Hoxby (2000), a researcher could impose a functional form based on Lazear (2001)'s model of achievement as a function of class size. Bajari, Hong, Park, and Town (2011) present a principal-agent model to study how insurers reimburse hospitals, where the marginal reimbursement rate is discontinuous on health expenditures.

This paper proposes consistent and asymptotically normal estimation procedures for weighted average treatment effects. The interpretation of the average effect and the estimation procedure depends on the case of heterogeneity assumptions. In the first case of heterogeneity assumptions, the average effect is a weighted average of local treatment effects (ALTE) computed at the finite number of cutoffs observed in the data. Such an ALTE is a meaningful summary measure for counterfactual scenarios that weight individuals close to the existing cutoffs. In the high school assignment example, suppose one town decides to marginally expand the capacity of some of its best high schools. Students that are currently near the admission cutoffs of such schools will be granted access to better school quality. The relative proportion of capacity increase across targeted high schools produces the relevant weighting scheme for the ALTE of this policy.

In the second case of heterogeneity assumptions, the treatment effects are described by a smooth function of cutoffs and doses. The average treatment effect (ATE) is a weighted integral of the treatment effect function $\beta(.)$ over the support set of variation in cutoffs and dose values. Such an ATE is a meaningful summary measure for counterfactual scenarios that weight individuals not only at the existing cutoffs but also between cutoffs. For example, suppose a new 'charter' school in a given town admits students by randomly drawing from a target population. This target population has a distribution of scores within the support of the observed distribution of scores in the data. The relevant ATE averages over the entire distribution of students that are granted access to this higher quality charter school, which includes scores between the existing cutoff values. Finally, in the third case of heterogeneity assumptions, interest lies not only in the ATE, which is the weighted integral of the para-

metric treatment effect function, but also on the parameters of this function. For example, a polynomial functional form on scores and average peer performance can be used to test the hypothesis that the returns of going to a better school is constant over scores.

The estimation procedure for the ALTEs and ATEs consists of two steps. The first step is identical in all three cases of heterogeneity assumptions. In the first step, we estimate the local treatment effects at each cutoff non-parametrically by local polynomial regression. The second step depends on the case of heterogeneity assumptions the researcher is willing to make. In the first case of heterogeneity assumptions, the second step estimate for the ALTE is a simple weighted average of first step estimates, where the researcher chooses the weighting scheme. In the second case of heterogeneity assumptions, the second step estimation consists of estimating the function $\beta(x, d, d')$ non-parametrically using a local polynomial regression of the first step estimates on cutoff values (x, d, d'). The estimator for the ATE is simply the weighted integral of the estimated function $\widehat{\beta}(x, d, d')$, where the researcher specifies the weighting density according to the relevant counterfactual scenario.

A key contribution of this paper is the inference on the ATE in RDD when the treatment effect function is unknown and of infinite dimension. Consistency for the integral ATE requires an asymptotic sequence where both the number of observations and cutoffs go to infinity. Our estimator for the ATE is asymptotically normal and its maximum convergence rate is root-n. Lastly, in the third case of heterogeneity assumptions, estimates for the parametric functional form are obtained by regressing the first step estimates on functions of cutoff-dose values specified by the researcher. Observations from different cutoffs are optimally weighted in this second step regression depending on the first step variance. The ATE in the third case of heterogeneity assumptions is the weighted integral of the estimated parametric functional form. We show consistency and asymptotic normality of all estimators.

Another advantage of the third case of heterogeneity assumptions is that a parametric functional form gives identification in the fuzzy RDD case. In the high school assignment example, when a student is accepted for a high school with peer-quality d, she may choose

to attend a different high school d' in the fuzzy case. When there are many cutoffs, each cutoff may exhibit its own compliance behavior. The observed average outcome of students around cutoff c is a weighted average of potential outcomes $Y_i(d)$ for the different school qualities d. Thus, comparing the academic performance of students just below the cutoff to those just above gives a mix of treatment effects of various doses.

To disentangle the different treatment effects, we make a minimal assumption about how the behavior of students changes with their test score by ruling out 'defiance'. If the test score of a student currently attending high school B increases so as to grant her access to high school A, we assume she either chooses to attend school A or stay at school B, and that she is not triggered to attend some other school C. When she chooses school A, we say she complies to the treatment eligibility change associated with the cutoff of school A. We call a student 'ever-complier' when she complies to the treatment change for at least one of the cutoffs and does not respond to the treatment change of other cutoffs. No-defiance is not a sufficient condition for the identification of treatment effects on ever-compliers in all cutoffs. For example, suppose we have a town with three schools and two cutoffs. At the second cutoff that grants admission to the best school, there could be ever-compliers that change from the worst school into the best school, and from the second best to the best school. We only observe the average change in the outcome variable aggregated over these two types of ever-compliers and cannot separately identify their treatment effects without further functional form assumptions. The parametric functional form and no-defiance are sufficient conditions for identification of treatment effects on ever-compliers. We provide a two-step estimator for such a parametric functional form and show consistency and asymptotic normality.

The remainder of this paper is divided into three main sections for each case of heterogeneity assumptions. Section 2 sets up the notation for RDD with multiple cutoffs to be used in all sections, describing the estimation procedure and providing sufficient conditions for inference on the ALTE in the first case of heterogeneity assumptions. Section 3 derives an estimator for the ATE and lays down sufficient conditions for valid inference in the sec-

ond case of heterogeneity assumptions. Section 4 explains the estimation procedure for the parameters of a functional form specified by the researcher in the third case of heterogeneity assumptions; sufficient conditions are given for the asymptotic inference in both the sharp and fuzzy RDD cases. Section 5 illustrates our methods using data from PEU. Section 6 concludes. The proofs are found in the Appendix.[4]

2 Average of Local Treatment Effects

In this section we assume the sharp RDD case and that the researcher believes that changes in treatment doses do not bear any quantitative relationship across cutoffs. Individuals are subject to different treatments across cutoffs, but the treatment dose cannot be summarized into a scalar variable. First, we define the notation to be used throughout the paper for the sharp RDD with multiple thresholds. Second, we define an average treatment effect (ATE) parameter where the researcher chooses the counterfactual distribution of interest. In this section, the choice of counterfactual distributions is limited to discrete distributions with support equal to the set of cutoffs and treatment doses in the data. In this case, the ATE is an average of local treatment effects (ALTE). The set of possible counterfactual distributions is expanded in the next sections. Lastly, we describe an estimation procedure for the ATE parameter and show consistency and asymptotic normality when the sample size grows large but the number of cutoffs is held fixed.

There are many cutoffs c defined on one scalar forcing variable X that assign individuals to different treatment doses defined by the variable D. Only in this section, the variable D is a qualitative measure of the treatment dose received. The assignment of individuals to treatments is observed for different sub-populations where the cutoffs and doses vary according to the sub-population. In the example of high school assignment, a sub-population is a town-year. Individual random variables are indexed by 'i'. 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

[4]The Appendix is available online at *www.stanford.edu/∼bertanha/Bertanha_JMP_appendix.pdf*

treatment doses is defined as $\mathcal{D}$. The discrete variable P_i takes values in $\mathcal{P} = \{1, \cdots, P\}$ and indicates the sub-population of individual 'i'. In sub-population $p \in \mathcal{P}$, a cutoff c is indexed by p, j where $j \in \mathcal{J}_p = \{1, \ldots, K(p)\}$ with $K(p)$ being the total number of cutoffs in sub-population p. The cutoffs are ordered such that $c_{1,p} < c_{2,p} < \ldots < c_{K(p),p}$. Sharp assignment means that an individual with forcing variable X_i in sub-population P_i is deterministically assigned to treatment dose $D_i = D(X_i, P_i)$ for some known assignment mapping $D : \mathcal{X} \times \mathcal{P} \to \mathcal{D}$ where

$$D(x,p) = \begin{cases} d_{p,0} & \text{if } c_{p,0} \leq x < c_{p,1} \\ d_{p,1} & \text{if } c_{p,1} \leq x < c_{p,2} \\ \vdots & \\ d_{p,K(p)} & \text{if } c_{p,K(p)} \leq x \leq c_{p,K(p)+1} \end{cases} \tag{1}$$

with $c_{0,p} = \underline{\mathcal{X}}$, $c_{K(p)+1,p} = \overline{\mathcal{X}}$. Each cutoff is characterized by three variables, $\mathbf{c}_{p,j} = (c_{p,j}, d_{p,j-1}, d_{p,j})$: the scalar threshold $c_{p,j}$; the treatment dose $d_{p,j-1}$ the individual receives if $c_{p,j-1} \leq X_i < c_{p,j}$; and the treatment dose $d_{p,j}$ the individual receives if $c_{p,j} \leq X_i < c_{p,j+1}$. The schedule of cutoffs and treatment doses is given by the non-random set[5] $\mathcal{C}_K = \{\mathbf{c}_{p,j}\}_{p \in \mathcal{P}, j \in \mathcal{J}_p}$ with $\mathcal{C}_K \subset \mathcal{C}$, where $\mathcal{C}$ is a subset of $\mathcal{X} \times \mathcal{D} \times \mathcal{D}$ of all possible cutoff and dose values. The total number of cutoffs from all sub-populations is $K = \sum_{p \in \mathcal{P}} K(p)$.

We follow the modern literature on treatment effects and use 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 can be summarized as follows. Values for the forcing variable X_i, sub-population P_i and potential outcomes $\{Y_i(d)\}_{d \in \mathcal{D}}$ are drawn iid $i = 1, ..., n$ from a joint distribution. Given the mapping $D(x, p)$, these n individuals are assigned to different treat-

[5]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.

ment doses $D_i = D(X_i, P_i)$. The observed outcome Y_i is given by

$$\begin{aligned} Y_i &= Y_i(D_i) = Y_i\left(D(X_i, P_i)\right) \\ &= \sum_{p \in \mathcal{P}} \sum_{j \in \mathcal{J}_p^0} Y_i(d_{p,j}) \mathbb{I}\{D_i = d_{p,j},\ P_i = p\} \end{aligned}$$

where $\mathcal{J}_p^0 = \mathcal{J}_p \cup \{0\} = \{0, 1, \ldots, K(p)\}$, and $\mathbb{I}\{\cdot\}$ is the indicator function.

The econometrician observes the schedule of cutoffs and treatment doses for all sub-populations and (Y_i, X_i, D_i, P_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 $d \to d'$. Consider a cutoff $\mathbf{c} = (c, d, d')$. RDD identifies the average treatment effect of changing d to d' over those individuals with forcing variable equal to c. Such local average treatment effect is denoted as

$$\beta(\mathbf{c}, p) \equiv \mathbb{E}[Y_i(d') - Y_i(d) | X_i = c, P_i = p] \tag{2}$$

The goal of this paper is to lay down conditions for combining multiple local average treatment effects at $\mathbf{c}$'s into a meaningful average treatment effect over set $\mathcal{C}$. To combine treatments effects from various sub-populations we make the following assumption.

Assumption 1. *For any* $\mathbf{c} = (c, d, d') \in \mathcal{C}$*, and any* $p \in \mathcal{P}$

$$\beta(\mathbf{c}, p) = \beta(\mathbf{c})$$

This 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. It does not restrict conditional means to be same across different sub-populations which accommodates time-trends and sub-population fixed effects for example.

It is widely known in the RDD literature that continuity of the conditional mean of

potential outcomes lead to identification of treatment effects at the cutoff values $\mathbf{c} \in \mathcal{C}_K$ (Hahn, Todd, and Van der Klaauw (2001)). This is re-stated in Lemma 1 for the multiple cutoff case.

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

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

There are two sources of heterogeneity for treatment effects across different cutoffs. Treatment effects are expected to be heterogeneous because of different values of the forcing-variable and changes in treatment doses. Researchers are often interested in summary measures of the different treatment effects. In this paper we work with average treatment effect as a default summary measure. A common practice in applied work is to normalize all cutoffs to zero in order to use existing estimation techniques for the one-cutoff case. This procedure produces an estimate for an average effect weighted by the relative density of the forcing variable at the cutoffs (Cattaneo, Keele, Titiunik, and Vazquez-Bare (2015)). An average treatment effect is only informative when the researcher deliberately chooses how different treatment effects are weighted. For example, a certain policy may have positive effects for some values of the forcing variable but negative effects for other values. Depending on how these effects are weighted, we may conclude the policy has no effect in a given population.

Existing data can be used to estimate an average treatment effect of the current policy but also new policies. The new policy applies a series of changes in treatment doses to a distribution of individuals. Each counterfactual policy scenario translates into a cumulative distribution function $F(\mathbf{c})$ for the forcing variable X and treatment dose changes (D, D') with support $\mathcal{C}$. Counterfactual distributions are restricted to $\mathcal{C}$ because that is the variation we observe in the data. Under certain assumptions on the distribution of unobserved heterogeneity, the average treatment effect of the counterfactual policy is the expected value

of the treatment effect function $\beta(\mathbf{c})$ under the counterfactual distribution $F(\mathbf{c})$.

$$\mu(F) = \mathbb{E}_F \beta(\mathbf{c}) \tag{3}$$

where $\mathbb{E}_F$ denotes the expectation under a distribution $F(\mathbf{c})$.

We explain the conditions under which (3) is the parameter of interest. Let the potential outcome be defined as a fixed response function of the observed characteristics (X_i, P_i) and an unobserved scalar variable U_i, that is, $Y_i(d) = \mathbb{Y}(X_i, d, P_i, U_i)$ for any $d \in \mathcal{D}$. Consider two worlds: (i) the world of the data generating process (dgp) with distribution $F^{dgp}_{X,D,P,U}$; and (ii) the counterfactual world where a new policy assigns individuals to changes in treatment doses. A counterfactual scenario is a complete distribution $F^*_{X,D,D',P,U}$ of forcing variables, treatment dose changes, and unobserved heterogeneity. Assuming that the response function $\mathbb{Y}(.)$ does not change between worlds (i) and (ii), the researcher is interested in estimating $\mathbb{E}_{F^*_{X,D,D',P,U}}[\mathbb{Y}(X_i, D_i', P_i, U_i) - \mathbb{Y}(X_i, D_i, P_i, U_i)]$. In practice, we only observe the marginal counterfactual distribution $F^*_{X,D,D',P}$. Moreover, the distribution of sub-population indices P_i does not matter because of assumption 1. Therefore, the counterfactual distribution we consider is the marginal $F^*_{X,D,D'} = F(\mathbf{c})$.

Lemma 2. *Assume that $F^{dgp}_{U|X,P} = F^*_{U|X,P} = F^*_{U|X,D,D',P}$, and that $\mathbb{Y}(.)$ does not change between the 'dgp' and 'counterfactual' worlds. Then, the average treatment effect of a policy with counterfactual distribution $F^*_{X,D,D',P,U}$ and observed marginal $F^*_{X,D,D'} = F$ is*

$$\mathbb{E}_{F^*_{X,D,D',P,U}}[\mathbb{Y}(X_i, D_i', P_i, U_i) - \mathbb{Y}(X_i, D_i, P_i, U_i)] = \mathbb{E}_F \beta(\mathbf{c}) = \mu(F)$$

The proof is a straightforward application of iterated expectations and is omitted. Lemma 2 states that the parameter $\mu(F)$ defined in equation (3) summarizes the average effect of a counterfactual policy F^* with observed marginal distribution F. At this level of generality, it is difficult to relax the assumptions of Lemma 2 without further knowledge of each specific application. Essentially, the conditions of Lemma 2 require the distribution of

unobservables conditional on observables to not change between 'dgp' and 'counterfactual' worlds. It also requires unobservables to be independent of the assignment of treatment in the 'counterfactual' world.

In this section, we consider the first case of heterogeneity assumptions. The first case covers those applications of RDD where treatment doses are not credibly summarized in a scalar treatment dose variable d. This restricts the set of counterfactual distributions to distributions with discrete support equal to $\mathcal{C}_K$, that is, the set of cutoff-dose values we observe in the data. For example, in the high school application, a new policy may marginally re-allocate students across the existing schools, and the parameter of interest may be the average of local average treatment effects at the existing cutoffs. The discrete counterfactual distribution F^{disc} is the probability mass function or weighting scheme $\{\omega_{p,j}\}_{p,j}$ over the set $\mathcal{C}_K = \{\mathbf{c}_{p,j}\}_{p,j}$. It represents the probability mass of students with test score equal to $c_{p,j}$ that undergo a change in school quality of $d_{p,j-1} \to d_{p,j}$ in this new policy. The average treatment effect $\mu(F^{disc})$ is identified under continuity of the conditional mean of potential outcomes (Lemma 1) for any choice of F^{disc} with support $\mathcal{C}_K$.

$$\mu(F^{disc}) = \sum_{p\in\mathcal{P}, j\in\mathcal{J}_p} \omega_{p,j}\ \beta(\mathbf{c}_{p,j})$$

The estimation of $\mu(F^{disc})$ is conducted in two steps. The first step uses observations near each of the cutoffs $c_{p,j}$ to non-parametrically estimate $B_{p,j}$ using local polynomial regression (LPR).

$$B_{p,j} = \lim_{e\downarrow 0} \{\mathbb{E}[Y_i|X_i = c_{p,j} + e, P_i = p] - \mathbb{E}[Y_i|X_i = c_{p,j} - e, P_i = p]\} \tag{4}$$

By Lemma 1, $B_{p,j} = \beta(\mathbf{c}_{p,j}) \equiv \beta_{p,j}$. The researcher chooses a bandwidth parameter $h_1 > 0$, a kernel density function $k(.)$, and the order of the polynomial regression $\rho_1 \in \mathbb{Z}_+$.[6]

[6]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 bandwidth parameter defines a neighborhood around each cutoff from which we use observations in the estimation. The farther observations are from the cutoff, the less weight they receive which is determined by the function $k(.)$. We fit a polynomial in X on each side of the cutoff, and the estimator $\hat{B}_{p,j}$ is the difference between the intercept of these two polynomial regressions.

$$\hat{B}_{p,j} = \hat{a}^{+}_{p,j} - \hat{a}^{-}_{p,j} \tag{5}$$

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

$$(\hat{a}^{-}_{p,j}, \hat{\mathbf{b}}^{-}_{p,j}) = \underset{(a,\mathbf{b})}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{X_i - c_{p,j}}{h_1}\right) v_i^{p,j-} \left[Y_i - a - b_1(X_i - c_{p,j}) - \ldots - b_{\rho_1}(X_i - c_{p,j})^{\rho_1}\right]^2 \tag{7}$$

where

$$v_i^{p,j+} = \mathbb{I}\{c_{p,j} \leq X_i < c_{p,j} + h_1,\ P_i = p\} \tag{8}$$

$$v_i^{p,j-} = \mathbb{I}\{c_{p,j} - h_1 < X_i < c_{p,j},\ P_i = p\} \tag{9}$$

The LPR estimator is known for low-order boundary bias and rate optimality (Fan and Gijbels (1996), Porter (2003)).

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

$$\hat{\mu}(F^{disc}) = \sum_{p\in\mathcal{P}, j\in\mathcal{J}_p} \omega_{p,j}\hat{B}_{p,j} \tag{10}$$

For the case of one sub-population, one cutoff, the asymptotic distribution of the LPR estimator $\hat{B}_{p,j}$ has been derived by Porter (2003). Minor adjustments give the distribution of each $\hat{B}_{p,j}$ and the weighted average. Asymptotic normality requires $n \to \infty$ and $h_1 \to 0$. Since the number of cutoffs K is fixed and $h_1 \to 0$, the neighborhoods around each cutoff don't overlap for large n. In large samples, each individual observation is used for only one $\hat{B}_{p,j}$ which makes $\hat{B}_{p,j} \perp \hat{B}_{p,l}$ for $j \neq l$. Therefore, the asymptotic distribution of $\hat{\mu}(F^{disc})$ will be the weighted sum of the asymptotic distributions of each $\hat{B}_{p,j}$. Below, we list sufficient conditions for the asymptotic normality result in Theorem 1.

Assumption 2. *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 3. *(i) For every $p \in \mathcal{P}$, the distribution $X_i|P_i = p$ has probability density function $f_{X|P}(x, p)$ and a bounded support $\mathcal{X} = [\underline{\mathcal{X}}, \overline{\mathcal{X}}]$; $f_{X|P}(x, p)$ is bounded and bounded away from zero uniformly on $(x, p) \in \mathcal{X} \times \mathcal{P}$.*

(ii) $f_{X|P}(x, p)$ is one time differentiable w.r.t x with partial derivative $\nabla_x f_{X|P}(x, p)$ bounded on $(x, p) \in \mathcal{X} \times \mathcal{P}$.

(iii) $\forall p \in \mathcal{P}$, $\mathbb{P}(P_i = p) = q_p > 0$

Assumption 4. *Let $\rho_1 \in \mathbb{Z}_+$ be the order of the first step LPR.*

(a) $R(x, d, p) = \mathbb{E}[Y_i(d)|X_i = x, P_i = p]$ is $\rho_1 + 1$ times continuously differentiable w.r.t. x with $\rho_1 + 1$-th partial derivative $\nabla_x^{\rho_1+1} R(x, d, p)$

(b) $\sigma^2(x,d,p) = \mathbb{E}\left\{\left[Y_i(d) - R(x,d,p)\right]^2 | X_i = x, P_i = p\right\}$ *is one time continuously differentiable w.r.t.* x *with partial derivative* $\nabla_x \sigma^2(x,d,p)$

Theorem 1. *Suppose assumptions 1, 2, 3, 4 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,*

$$\sqrt{nh_1}\left(\hat{\mu}(F^{disc}) - \mu(F^{disc})\right) \xrightarrow{d} N\left(C\sum_{p,j}\omega_{p,j}\mathcal{B}_{p,j};\ \sum_{p,j}\omega_{p,j}^2\mathcal{V}_{p,j}\right)$$

where

$$\mathcal{B}_{p,j} = \frac{1}{(\rho_1+1)!}\left[\nabla_x^{\rho_1+1} m(c_{p,j}^+, p) - (-1)^{\rho_1+1}\nabla_x^{\rho_1+1} m(c_{p,j}^-, p)\right] e_1'\Gamma^{-1}\gamma^* \tag{11}$$

$$\mathcal{V}_{p,j} = \frac{\zeta^2(c_{p,j}^+, p) + \zeta^2(c_{p,j}^-, p)}{f_{X|P}(c_{p,j}, p) q_p} e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1 \tag{12}$$

$$\nabla_x^{\rho_1+1} m(c_{p,j}^+, p) = \lim_{x \downarrow c_{p,j}} \nabla_x^{\rho_1+1}\mathbb{E}[Y_i | X_i = x, P_i = p]$$

$$\nabla_x^{\rho_1+1} m(c_{p,j}^-, p) = \lim_{x \uparrow c_{p,j}} \nabla_x^{\rho_1+1}\mathbb{E}[Y_i | X_i = x, P_i = p]$$

$$\zeta^2(c_{p,j}^+, p) = \lim_{x \downarrow c_{p,j}} \mathbb{E}\left\{\left(Y_i - \mathbb{E}[Y_i|X_i, P_i]\right)^2 \ \middle|\ X_i = x, P_i = p\right\}$$

$$\zeta^2(c_{p,j}^-, p) = \lim_{x \uparrow c_{p,j}} \mathbb{E}\left\{\left(Y_i - \mathbb{E}[Y_i|X_i, P_i]\right)^2 \ \middle|\ X_i = x, P_i = p\right\}$$

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

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

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

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

$$\rho_1 \text{ is the order of the LPR}$$

To perform inference using this asymptotic result, we need consistent estimators for the asymptotic bias and variance in equations 11 and 12. 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 $\nabla_x^{\rho_1+1} m(c_{p,j}^{\pm}, p)$, $\zeta^2(c_{p,j}^{\pm}, p)$, and $f_{X|P}(c_{p,j}, p)q_p$ which is a straightforward non-parametric problem. For the side derivatives $\nabla_x^{\rho_1+1} m(c_{p,j}^{\pm}, p)$, a consistent estimator is obtained from a LPR of order $\rho_1 + 1$ (equations 6 and 7) that uses observations from each side of the cutoff $c_{p,j}$. The estimator is simply the slope coefficient on $(x - c_{p,j})^{\rho_1+1}$. Lemma 8(scalar case) in the Appendix shows consistency of this estimator. The density $f_{X|P}(c_{p,j}, p)q_p$ is consistently estimated (Silverman (1986)) by

$$\widehat{f}_{X,P}(c_{p,j}, p) = \frac{1}{nh_1} \sum_{i=1}^{n} k\left(\frac{X_i - c_{p,j}}{h_1}\right) \mathbb{I}\{P_i = p\}$$

Porter (2003) suggests the following consistent estimation procedure for the side limits of the variance $\zeta^2(c_{p,j}^{\pm}, p)$:

$$\widehat{m}(x, p) = \frac{\frac{1}{nh_1} \sum_{i=1}^{n} k\left(\frac{X_i - c_{p,j}}{h_1}\right) \mathbb{I}\{P_i = p\} \left(Y_i - \sum_{j \in \mathcal{J}_p} \mathbb{I}\{c_{p,j} \leq x\} \widehat{B}_{p,j}\right)}{\widehat{f}_{X,P}(c_{p,j}, p)}$$

$$\widehat{\varepsilon}_i = Y_i - \widehat{m}(X_i, P_i) - \sum_{p \in \mathcal{P}, j \in \mathcal{J}_p} \mathbb{I}\{P_i = p, c_{p,j} \leq x\} \widehat{B}_{p,j}$$

$$\widehat{\zeta}^2(c^+_{p,j},p) = \frac{\frac{1}{nh_1}\sum_{i=1}^n v_i^{p,j+} k\left(\frac{X_i - c_{p,j}}{h_1}\right)\widehat{\varepsilon}_i^2}{\frac{1}{2}\widehat{f}_{X,P}(c_{p,j},p)}$$
$$\widehat{\zeta}^2(c^-_{p,j},p) = \frac{\frac{1}{nh_1}\sum_{i=1}^n v_i^{p,j-} k\left(\frac{X_i - c_{p,j}}{h_1}\right)\widehat{\varepsilon}_i^2}{\frac{1}{2}\widehat{f}_{X,P}(c_{p,j},p)}$$

Using these estimators along with equations 11 and 12 and definitions of Theorem 1, we obtain estimators for the asymptotic variance and bias of $\hat{\mu}(F^{disc})$. Alternative approaches to build confidence intervals include bootstrapping (Hardle and Bowman (1988), Neumann and Polzehl (1998)), the empirical likelihood methods of Otsu, Xu, and Matsushita (2014), or the robust confidence intervals of Calonico, Cattaneo, and Titiunik (2014).

3 General Average Treatment Effects

In the second case of heterogeneity assumptions, we assume the sharp RDD case and that treatment doses are credibly summarized in a scalar variable d. The treatment effect function $\beta(\mathbf{c})$ is assumed to be a smooth function of $\mathbf{c} = (c, d, d')$. In the high school assignment application, the treatment dose is the quality of each school. Examples of measures of school quality include the average test score of peers, the average number of teachers or funding per student. Using the variation of cutoff-doses across sub-populations, these heterogeneity assumptions allows for inference on an ATE computed over the entire support $\mathcal{C}$. In this section, we propose a two-step estimation procedure for such ATEs. We show consistency and asymptotic normality when the sample size and the number of cutoffs grow large.

In the first case of heterogeneity assumptions, the set of counterfactual distributions is limited to the set of discrete distributions with support $\mathcal{C}_K$. It is unlikely that new policies have the same schedule of cutoffs and treatment doses as the policy we observe in the data. Thus, the ability to compute weighted effects for cutoff and dose values beyond the existing ones becomes crucial. For example, we could infer the average treatment effect of giving Medicaid benefits to an entire neighborhood of individuals; or the average treatment

effect on all the students admitted into a new charter school. We consider policy questions that lead to a continuous counterfactual distribution F^{cont} with probability density function $\omega : \mathcal{C} \to \mathbb{R}$.[7] 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 such that cutoff-dose values are dense in the set $\mathcal{C}$ identifies $\mu(F^{cont})$.

Lemma 3. *Assume that (i) $\omega(\mathbf{c})$, $\beta(\mathbf{c})$ are continuous functions in the compact set $\mathcal{C} \subset \mathbb{R}^3$; and (ii) an infinite amount of data means that there is an infinite countable 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. Just as in section 2, in the first step, LPRs produce estimates for each $B_{p,j}$ using observations in the neighborhood of each cutoff p, j in the data (equations 5, 6 and 7). In the second step, we average the first step estimates using 'corrected weights' $\{\Delta_{p,j}\}_{p,j}$.

$$\widehat{\mu}(F^{cont}) = \sum_{p \in \mathcal{P}, j \in \mathcal{J}_p} \Delta_{p,j} \hat{B}_{p,j}$$

The 'corrected weights' are implicitly defined by integrating a non-parametric estimate of $\beta(\mathbf{c})$ weighted by $\omega(\mathbf{c})$. There is a closed form expression for $\{\Delta_{p,j}\}_{p,j}$ in equation (15) below. At any point $\mathbf{c}$, the function $\beta(\mathbf{c})$ is non-parametrically estimated through a multivariate local polynomial regression of $\widehat{B}_{p,j}$ on $\mathbf{c}_{p,j}$ in a h_2-neighborhood of point $\mathbf{c}$. The researcher chooses a bandwidth $h_2 > 0$, a polynomial degree $\rho_2 \geq 3$, and a kernel density function $k(.)$. For each point $\mathbf{c} \in \mathcal{C}$, the estimate $\widehat{\beta}(\mathbf{c})$ is obtained by the following least squares

[7] We choose to focus on the more interesting case of continuous counterfactual distributions, but the same results with minor changes apply to discrete or mixed counterfactual distributions.

minimization:

$$\widehat{\beta}(\mathbf{c}) = \widehat{\eta}_1 \tag{13}$$

$$\widehat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta}}{\operatorname{argmin}} \left(\widehat{\mathbf{B}} - \mathbf{E}(\mathbf{c})\boldsymbol{\eta}\right)' \Omega_{h_2}(\mathbf{c}) \left(\widehat{\mathbf{B}} - \mathbf{E}(\mathbf{c})\boldsymbol{\eta}\right) \tag{14}$$

where $\mathbf{B}$ is the $K \times 1$ vector formed by stacking all of the $B_{p,j}$; $\mathbf{E}(\mathbf{c})$ is the $K \times J$ matrix formed by stacking the K, $J \times 1$ vectors, $E_{p,j}(\mathbf{c})$; $E_{p,j}(\mathbf{c})$ is a vector of the standard basis of the space of polynomials in $\mathbb{R}^3$ of maximum degree ρ_2 evaluated at $\mathbf{c} - \mathbf{c}_{p,j}$; each entry in $E_{p,j}(\mathbf{c})$ is a polynomial of the form $(\mathbf{c} - \mathbf{c}_{p,j})^{\boldsymbol{\gamma}}$ where $\boldsymbol{\gamma} \in \mathbb{Z}^3_+$, $|\boldsymbol{\gamma}| \equiv \gamma_1 + \gamma_2 + \gamma_3 \leq \rho_2$, $(\mathbf{c} - \mathbf{c}_{p,j})^{\boldsymbol{\gamma}} = (c - c_{p,j})^{\gamma_1}(d - d_{p,j-1})^{\gamma_2}(d' - d_{p,j})^{\gamma_3}$; the first entry in $E_{p,j}(\mathbf{c})$ is the polynomial of degree zero (i.e. 1), the next 3 entries are all the polynomials of degree 1 (i.e. all $(\mathbf{c} - \mathbf{c}_{p,j})^{\boldsymbol{\gamma}}$ such that $|\boldsymbol{\gamma}| = 1$), and then all the polynomials of degree 2, and so on, until degree ρ_2; therefore, $J = \binom{\rho_2+3}{3}$; $\boldsymbol{\eta}$ is a $J \times 1$ vector of parameters, and η_1 is the first coordinate of the vector $\boldsymbol{\eta}$; $\Omega_{h_2}(\mathbf{c})$ is the $K \times K$ diagonal matrix of kernel weights:

$$\Omega_{h_2}(\mathbf{c}) = diag\left\{\Omega_{h_2,p,j}\right\}_{p,j} = diag\left\{k\left(\frac{x - c_{p,j}}{h_2}\right) k\left(\frac{d - d_{p,j-1}}{h_2}\right) k\left(\frac{d' - d_{p,j}}{h_2}\right)\right\}_{p,j}$$

The estimator $\widehat{\mu}(F^{cont})$ is the weighted integral of $\widehat{\beta}(\mathbf{c})$ over set $\mathcal{C}$ with weighting density $\omega(\mathbf{c})$ chosen by the researcher. Alternatively, the average treatment effect estimator is simply a weighted sum of first stage estimates $\hat{B}_{p,j}$.

$$\widehat{\mu}(F^{cont}) = \int_{\mathcal{C}} \omega(\mathbf{c})\widehat{\beta}(\mathbf{c})\ d(\mathbf{c}) = \sum_{p \in \mathcal{P}, j \in \mathcal{J}_p} \Delta_{p,j}\hat{B}_{p,j}$$

where $\Delta_{p,j}$ are the 'corrected weights' given by the formula

$$\Delta_{p',j'} = \int_{\mathcal{C}} \omega(\mathbf{c})\ e_1' \left[\mathbf{E}(\mathbf{c})'\Omega_{h_2}(\mathbf{c})\mathbf{E}(\mathbf{c})\right]^{-1} \Omega_{h_2,p',j'}(\mathbf{c})\mathbf{E}_{p',j'}(\mathbf{c})\ \ d(\mathbf{c})$$

$$= \int_{\mathcal{C}} \omega(\mathbf{c}) \frac{det\left(\mathbf{E}(\mathbf{c})'\Omega_{h_2}(\mathbf{c})\mathbf{E}_{\mathbf{0}\leftarrow e_{p',j'}}(\mathbf{c})\right)}{det\left(\mathbf{E}(\mathbf{c})'\Omega_{h_2}(\mathbf{c})\mathbf{E}(\mathbf{c})\right)} \; d(\mathbf{c}) \tag{15}$$

where e_1 is the $J \times 1$ vector $e_1 = [1 \quad 0 \quad 0 \quad \cdots \quad 0]'$, and $\mathbf{E}_{\mathbf{0}\leftarrow e_{p',j'}}(\mathbf{c})$ is the matrix valued function $\mathbf{E}(\mathbf{c})$ except for the first column which is replaced by the $K \times 1$ vector $e_{p',j'}$ that is zero everywhere except for the (p', j')-th entry which is equal to 1.

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 $K \to \infty$. Our asymptotic exercise has the sample size $n \to \infty$, the total number of cutoffs $K \to \infty$, but the number of sub-populations P is fixed.[8] The integral of a function can be approximated by the 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. In our case, the function evaluations are estimated which means that the integral approximation error has to converge to zero fast enough to ensure asymptotic normality.

Assumption 5 states conditions on the limiting behavior of the schedule of cutoffs and treatment doses and on how it approximates set $\mathcal{C}$.

Assumption 5.

(a) The schedule of cutoffs and doses comes from a triangular array of fixed constants that depends on K (total number of cutoffs) denoted $\mathcal{C}_K$:

$$\mathcal{C}_K = \{\mathbf{c}_{p,j,K}\}_{p\in\mathcal{P}, j\in\mathcal{J}_{p,K}}$$

where $\mathcal{C}_K \subset \mathcal{C}$ and $\mathcal{J}_{p,K} = \{1, \ldots, K_K(p)\}$ with total number of cutoffs in subpopulation p indexed by K;

[8] Another less tractable asymptotic exercise could have $n \to \infty$, $P \to \infty$ but a fixed number of cutoffs $K(p)$ in each sub-population. This alternative asymptotics has the probability of some sub-populations converge to zero. Conditional mean estimators have these probabilities in the denominator which complicates their asymptotics, and we chose not to pursue this route. Moreover, our pooling assumption 1 makes an additional treatment effect in an existing sub-population indistinguishable from the same treatment effect in an additional sub-population.

(b) The maximum distance between two consecutive cutoffs within a subpopulation $p \in \mathcal{P}$ is inversely proportional to the total number of cutoffs $K(p)$ in that subpopulation

$$\max_{j=1,\ldots,K_K(p)} |c_{p,j,K} - c_{p,j-1,K}| = O\left(K_K(p)^{-1}\right)$$

(c) given the second step estimation bandwidth sequence $h_{2,K}$, the number of points in $\mathcal{C}_K$ that are within the $h_{2,K}$ neighborhood of any point in $\mathcal{C}$ is of order $Kh_{2,K}^3$:

$$\sup_{\mathbf{c}\in\mathcal{C}} \sum_{j=1}^{K} \mathbb{I}\left\{\Omega_{h_{2,K},j}(\mathbf{c}) > 0\right\} = O\left(Kh_{2,K}^3\right)$$

where $\Omega_{h_{2,K},j}(\mathbf{c})$ is defined after equation 14;

(d) for some $\rho_2 \geq 3$ (degree of 2nd step local polynomial)

$$\sup_{\substack{\mathbf{c}\in\mathcal{C} \\ 1\leq j\leq K}} \left| \frac{det\left(\mathbf{E}(\mathbf{c})'\Omega_{h_{2,K}}(\mathbf{c})\mathbf{E}_{\mathbf{0}\leftarrow e_j}(\mathbf{c})\right)}{det\left(\mathbf{E}(\mathbf{c})'\Omega_{h_{2,K}}(\mathbf{c})\mathbf{E}(\mathbf{c})\right)} \right| = O\left((Kh_{2,K}^3)^{-1}\right)$$

where $\mathbf{E}(\mathbf{c})$ and $\Omega_{h_{2,K}}(\mathbf{c})$ were defined after equation 14; $\mathbf{E}_{\mathbf{0}\leftarrow e_j}(\mathbf{c})$ is equal to $\mathbf{E}(\mathbf{c})$ except for the first column which has the $K \times 1$ vector e_j which is zero everywhere except for j-th coordinate that is equal to 1.

Conditions in assumption 5 basically require two things: for large K, (i) the proportion of observations in the h_2 neighborhood of any point in set $\mathcal{C}$ is of the same order of the volume of this neighborhood, that is, h_2^3; (ii) there is always enough points in the h_2 neighborhood of any point in set $\mathcal{C}$ for the invertibility of the $(\mathbf{E}(\mathbf{c})'\Omega_{h_2}(\mathbf{c})\mathbf{E}(\mathbf{c}))$ matrix. These conditions are satisfied in a variety of examples of triangular arrays of points, and we give one simple example in the Appendix (section 7.9.3) where these conditions hold. Next, we state sufficient smoothness conditions on the functions $\omega(\mathbf{c})$ and $\beta(\mathbf{c})$.

Assumption 6. *(a) $\omega(\mathbf{c})$ is a continuous function of $\mathbf{c}$;*

(b) for $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3) \in \mathbb{Z}_+^3$, *let*

$$\nabla^{|\gamma|}\beta(\mathbf{c}) = \frac{\partial^{|\gamma|}\beta(\mathbf{c})}{\partial \mathbf{c}^{\boldsymbol{\gamma}}}$$

denote the partial derivatives of $\beta(\mathbf{c})$ *with respect to* $\mathbf{c}$ *where* $\mathbf{c}^{\boldsymbol{\gamma}} = c^{\gamma_1} d^{\gamma_2} d'^{\gamma_3}$ *and* $|\boldsymbol{\gamma}| = \gamma_1 + \gamma_2 + \gamma_3$. *Assume* $\nabla^{|\gamma|}\beta(\mathbf{c})$ *is continuous for every* $\boldsymbol{\gamma}$ *such that* $|\boldsymbol{\gamma}| \leq \rho_2 + 1$, *where* ρ_2 *is the polynomial degree in the second step estimation.*

(c) $R(x, d, p) \equiv \mathbb{E}[Y_i(d)|X_i = x, P_i = p]$ *is* $\rho_1 + 2$ *times continuous differentiable wrt* x *with* $\rho_1 + 2$*-th partial derivative* $\nabla_x^{\rho_1+2} R(x, d, p)$ *where* ρ_1 *is the order of the LPR in the first step estimation*

(d) $\nabla_x^{\rho_1+2} R(x, d, p)$ *and* $\nabla_x \sigma^2(x, d, p)$ *are bounded functions of* (x, d, p)

(e) $Y_i(d) - R(x, d, p)$ *is a bounded random function of* (x, d, p) *a.s.*

(f) for the schedule of cutoffs and doses $\mathcal{C}_K$ *and corrected weights* $\Delta_{p,j,K}$ *(defined in equation 15) the limits below are well defined*

$$\lim_{K\to\infty} \left\{ \sum_{p,j} \Delta_{p,j,K} \left[\nabla_x^{\rho_1+1} R(c_{p,j,K}, d_{p,j,K}, p) - (-1)^{\rho_1+1} \nabla_x^{\rho_1+1} R(c_{p,j,K}, d_{p,j-1,K}, p) \right] \right\}$$

$$\lim_{K\to\infty} \left\{ K \sum_{p,j} \Delta_{p,j,K}^2 \frac{\sigma^2(c_{p,j,K}, d_{p,j,K}, p) + \sigma^2(c_{p,j,K}, d_{p,j-1,K}, p)}{f_{X|P}(c_{p,j,K}, p) q_p} \right\}$$

Theorem 2 states the rate conditions under which our estimator $\widehat{\mu}(F^{cont})$ has an asymptotic normal distribution.

Theorem 2. *Assume conditions in 1, 2, 3, 4, 5, 6 hold.*

As $n \to \infty$, *assume that* $K \to \infty$, $h_1 \to 0$, *and* $h_2 \to 0$ *such that*

- $\sqrt{Knh_1} h_1^{\rho_1+1} \to C \in [0, \infty)$ *where* ρ_1 *is the order of the first step LPR*

- $\frac{\sqrt{K}\log n}{\sqrt{nh_1}} \to 0$ *and* $Kh_1 = O(1)$
- $\sqrt{Knh_1}h_2^{\rho_2+1} \to 0$ *and* $1/Kh_2^3 = O(1)$ *where* ρ_2 *is the order of the second step multivariate LPR.*

then

$$\sqrt{Knh_1}\left(\widehat{\mu}(F^{cont}) - \mu(F^{cont})\right) \xrightarrow{d} N\left(C \cdot AB; AV\right)$$

where

$$AB = \lim_{K,n\to\infty}\left\{\sum_{p,j}\Delta_{p,j}\mathcal{B}_{p,j}\right\}$$

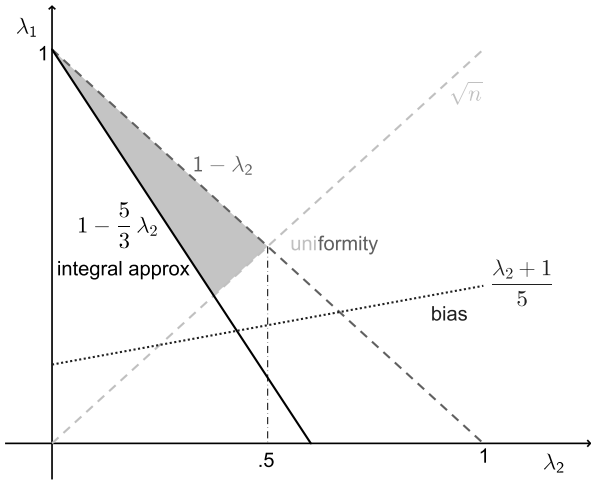
$$AV = \lim_{K,n\to\infty}\left\{K\sum_{p,j}\Delta_{p,j}^2\mathcal{V}_{p,j}\right\}$$

and $\mathcal{B}_{p,j}$, $\mathcal{V}_{p,j}$, $\Delta_{p,j}$ *are given in equations (11), (12) and (15).*

We give a simple example to illustrate 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 $h_2 = MK^{-1/3}$ for some constant M. First, note that $h_2 \to 0$ and $Kh_2^3 = O(1)$. 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 set of rate conditions 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 (dotted lines); 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 each of the cutoffs to insure the uniformity results (dashed lines). The third rate condition is equivalent to $1 + \lambda_2\left(1 - \frac{2}{3}(\rho_2 + 1)\right) < \lambda_1$ or $1 - \lambda_2\left(\frac{5}{3}\right) < \lambda_1$: K has to grow fast enough relatively to n to insure the integral approximation error vanishes faster than the rate of convergence of the estimator $\widehat{\mu}(F^{cont})$ (solid line). 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 with $\rho_2 = 3$ we do not have asymptotic bias since the shaded area does not touch the bias dotted line. We can use a higher second step polynomial degree of at least $\rho_2 = 6$ (expands the shaded area to the left) allowing combinations of λ_1 and λ_2 that lead to asymptotic bias. The maximum convergence rate of the estimator $\sqrt{Knh_1}$ is equal to $\sqrt{n}$ along the dashed line labeled $\sqrt{n}$.

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 $h_2 = K^{-1/3}$. 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: $\sqrt{Knh_1}h_1^{\rho_1+1} \to C$ (dotted line); $\sqrt{K}\log n/\sqrt{nh_1} \to 0$ and $Kh_1 = O(1)$ (dashed lines); and $\sqrt{Knh_1}h_2{}^{\rho_2+1} \to 0$ (solid line).

Another point worth mentioning is that the finite sample bias and standard error expressions are the same for both finite or infinite K asymptotics as long as we use corrected weights. If we use corrected weights to construct a discrete counterfactual distribution F^{disc}, we have

$$\mu(F^{disc}) = \textstyle\sum_{p,j} \Delta_{p,j}\beta_{p,j}$$

Then, the confidence interval for $\mu(F^{disc})$ (asymptotics with finite K) has the same formula as the confidence interval for $\mu(F^{cont})$ (asymptotics with infinite K):

$$CI(\mu(F^{disc});\ 1-\alpha) = CI(\mu(F^{cont});\ 1-\alpha)$$
$$= \left[\widehat{\mu}(F^{cont}) - h_1^{\rho_1+1}\sum_{p,j}\Delta_{p,j}\mathcal{B}_{p,j} \pm z_{\alpha/2}\sqrt{\sum_{p,j}\Delta_{p,j}^2\mathcal{V}_{p,j}}\right]$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution. This confidence interval has asymptotic coverage of $1-\alpha$ under both types of asymptotics (finite or infinite K).

The asymptotic bias and variance terms of Theorem 2 can be consistently estimated using the procedures discussed at the end of section 2. One has to compute $\widehat{\mathcal{B}_{p,j}}$ and $\widehat{\mathcal{V}_{p,j}}$ and the weighted sum using the corrected weights $\Delta_{p,j}$. Sufficient moment and rate conditions give consistency of these estimators under the asymptotics with large number of cutoffs.

4 Parametric Heterogeneity

In the third case of heterogeneity assumptions, the researcher specifies a parametric functional form for the treatment effect function $\beta(\mathbf{c})$. Economic theory or a priori knowledge guides the choice of a functional form that credibly summarizes the heterogeneity of treatment effects. For example, Lazear (2001) presents a well known formalization of a theory of educational output as a function of class size, teacher quality and student characteristics. In this section we discuss both the sharp and fuzzy RDD cases in separate subsections. For each case, we give sufficient conditions for identification and asymptotically normal estimation of the functional form parameters. The ATE is simply a linear combination of these parameters. Different than section 3, identification does not require a large number of cutoffs because of the parametric functional form assumption. The default asymptotic exercise has the sample size growing to infinity but the number of cutoffs fixed. In the sharp case, we show that our asymptotic normality result also holds when the number of cutoffs goes to

infinity.

4.1 Sharp RDD

Thus far we considered ATEs over general distributions of individuals F when the treatment effect function $\beta(\mathbf{c})$ is an unknown 'infinite' dimensional object. In this section, economic theory or a priori knowledge of the researcher restricts $\beta(\mathbf{c})$ to be an unknown 'finite' dimensional object. Besides the ATE, researchers are also interested in learning about the parameters of $\beta(\mathbf{c})$ and how treatment effects vary with changes in $\mathbf{c}$. For example, in the high school assignment application, we may be interested in learning whether the average return to school quality varies with the test score which is a measure of ability. If $\beta(\mathbf{c};\boldsymbol{\theta}) = \theta_1(d'-d) + \theta_2(d'-d)x + \theta_3(d'-d)x^2$ for unknown $\boldsymbol{\theta}$, we can test the hypothesis that $\theta_3 = 0$. The parametric assumption is stated as following.

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 $\mathbb{E}_F[\boldsymbol{\mathcal{W}}(c,d') - \boldsymbol{\mathcal{W}}(c,d)]$ is well defined for the counterfactual distribution F. The treatment effect function $\beta(\mathbf{c})$ is equal to the known vector valued function $\boldsymbol{\mathcal{W}}(c,d') - \boldsymbol{\mathcal{W}}(c,d)$ times an unknown vector of paramaters $\boldsymbol{\theta}_0 \in \mathbb{R}^q$:*

$$\beta(c,d,d') = [\boldsymbol{\mathcal{W}}(c,d') - \boldsymbol{\mathcal{W}}(c,d)]'\boldsymbol{\theta}_0$$

It is worth mentioning that assumption 7 is weaker than a a parametric functional form on the conditional mean function of the outcome Y_i, a common practice in applied work. A parametric assumption on $\beta(\mathbf{c})$ is equivalent to a semi-parametric assumption on the conditional mean of $\mathbb{E}[Y_i(d)|X_i = c, P_i = p]$. To see this, fix a baseline treatment dose $d_0 \in \mathcal{D}$. For any $(c,d,p) \in \mathcal{X}\times\mathcal{D}\times\mathcal{P}$,

$$\mathbb{E}[Y_i(d)|X_i = c, P_i = p] = \beta(c,d_0,d) + \mathbb{E}[Y_i(d_0)|X_i = c, P_i = p]$$

Under assumption 7, if you know $\boldsymbol{\theta}_0$, you know the entire function $\beta(\mathbf{c})$, but you still need knowledge of $\mathbb{E}[Y_i(d_0)|X_i = c, P_i = p]$ as a function of (c, p) in order to retrieve the entire function $\mathbb{E}[Y_i(d)|X_i = c, P_i = p]$ for all (c, d, p). In other words, our functional form restriction is robust to misspecification of $\mathbb{E}[Y_i(d_0)|X_i = c, P_i = p]$. Robustness to misspecification in the conditional mean of Y_i is an useful property because empirical evidence suggests the conditional mean of Y_i to be a much more complex function than the treatment effect function $\beta(\mathbf{c})$. In this case, misspecifying the conditional mean of Y_i leads to a larger bias than misspecifying the treatment effect function $\beta(\mathbf{c})$.

The ATE in the third case of heterogeneity assumptions is simply the expectation of the functional form of assumption 7 with under the counterfactual distribution F chosen by the researcher.

$$\begin{aligned}\mu(F) &= \mathbb{E}_F\left[\beta(\mathbf{c};\boldsymbol{\theta}_0)\right] \\ &= \mathbb{E}_F\left[\boldsymbol{\mathcal{W}}(c,d') - \boldsymbol{\mathcal{W}}(c,d)\right]'\boldsymbol{\theta_0} \equiv \mathbf{Z}(F)\boldsymbol{\theta}_0 \end{aligned} \tag{16}$$

where $\mathbf{Z}$ is a known $1 \times q$ vector.

$$\mathbf{Z}(F) \equiv \mathbb{E}_F\left[\boldsymbol{\mathcal{W}}(c,d') - \boldsymbol{\mathcal{W}}(c,d)\right]' \tag{17}$$

Lemma 4 shows that θ_0 and $\mu(F)$ are identified as long as there is sufficient variation in cutoff characteristics relative to the basis functions $[\boldsymbol{\mathcal{W}}(c,d') - \boldsymbol{\mathcal{W}}(c,d)]$.

Lemma 4. *Suppose assumptions 1 and 7 hold, and that $\beta(\mathbf{c})$ is identified for $\forall \mathbf{c} \in \mathcal{C}_K$. Let $W_{p,j} = \boldsymbol{\mathcal{W}}(c_{p,j}, d_{p,j}) - \boldsymbol{\mathcal{W}}(c_{p,j}, d_{p,j-1})$, where $\boldsymbol{\mathcal{W}}(c,d)$ is the $q \times 1$ vector valued function defined in assumption 7, and $\mathbf{W}$ is the $K \times q$ matrix formed by stacking $W_{p,j}$ for all p, j. If $\mathbf{W}'\mathbf{W}$ is invertible, then $\boldsymbol{\theta}_0$ is identified and equal to $(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{B}$, where $\mathbf{B}$ is the stacking of $B_{p,j}$ for all p, j (equation 4).*

The estimation of $\boldsymbol{\theta}_0$ is conducted in two steps. The first step, we use observations

near each of the cutoffs in each sub-population to estimate $B_{p,j}$ non-parametrically by LPR (equations 5, 6 and 7). In the second stage, we regress $\hat{B}_{p,j}$ on the basis functions evaluated at each cutoff-dose values $W_{p,j}$ to obtain $\widehat{\boldsymbol{\theta}}$. Since the treatment effect function is parametric, we can weight first stage estimates differently to minimize the mean squared error (MSE) of $\widehat{\boldsymbol{\theta}}$. More specifically, using a $K \times K$ symmetric and positive definite weighting matrix Ω chosen by the researcher, $\widehat{\boldsymbol{\theta}}$ is the solution to the following weighted least squares problem:

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

As in equation 16, the estimator for $\mu(F)$ is a linear combination of $\widehat{\boldsymbol{\theta}}$.

$$\widehat{\mu}(F) = \mathbf{Z}(F)\widehat{\boldsymbol{\theta}}$$

For a fixed number of cutoffs, as the sample size n increases and the bandwidth h_1 converges to zero, each individual observation is used only once in the whole estimation after a large n. The estimated treatment effects are independent of each other across different cutoffs. The asymptotic distribution of each element of $\widehat{\boldsymbol{\theta}}$ is a linear combination of the asymptotic normal distribution of each $\hat{B}_{p,j}$ (Lemma 8(scalar case) in the Appendix)

Theorem 3. *Suppose assumptions 1, 2, 3, 4, 7 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,*

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

$$\sqrt{nh_1}\left(\widehat{\mu}(F) - \mu(F)\right) \xrightarrow{d} N\Big(C\mathbf{Z}(F)(\mathbf{W}'\Omega\mathbf{W})^{-1}\mathbf{W}'\Omega\mathcal{B};$$

$$\mathbf{Z}(F)(\mathbf{W}'\Omega\mathbf{W})^{-1}\mathbf{W}'\Omega\mathcal{V}\Omega\mathbf{W}(\mathbf{W}'\Omega\mathbf{W})^{-1}\mathbf{Z}(F)'\Big)$$

Moreover, the asymptotic MSE of either $\sqrt{nh_1}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)$ *or* $\sqrt{nh_1}\left(\widehat{\mu}(F)-\mu(F)\right)$ *is minimized when* $\Omega=(C^2\mathcal{B}\mathcal{B}'+\mathcal{V})^{-1}$. *Below, the definitions used:*

$$\mathcal{B}=\left[\mathcal{B}_{1,K(1)},\ldots,\mathcal{B}_{K(1),K(1)},\mathcal{B}_{1,K(2)},\ldots,\mathcal{B}_{K(P),K(P)}\right]'$$

the formula for $\mathcal{B}_{p,j}$ *is given in equation 11*

$$\mathcal{V}=diag\left\{\mathcal{V}_{1,K(1)},\ldots,\mathcal{V}_{K(1),K(1)},\mathcal{V}_{1,K(2)},\ldots,\mathcal{V}_{K(P),K(P)}\right\}'$$

the formula for $\mathcal{V}_{p,j}$ *is given in equation 12*

$\mathbf{Z}(F)$ *is defined in equation 17*

$\mathbf{W}$ *is defined in Lemma 4*

Estimates for $\hat{C}$, $\hat{\mathcal{B}}$, and $\hat{\mathcal{V}}$ along with knowledge of $\mathbf{W}$, $\mathbf{Z}(F)$ give the estimates for the asymptotic variance and bias. We obtain estimates for $\mathcal{B}_{p,j}$ and $\mathcal{V}_{p,j}$ according to the procedure discussed at the end of section 2. Once we have $\hat{\mathcal{B}}$ and $\hat{\mathcal{V}}$, we compute the optimal weighting matrix $\hat{\Omega}=(\hat{C}^2\hat{\mathcal{B}}\hat{\mathcal{B}}'+\hat{\mathcal{V}})^{-1}$, where $\hat{C}=\sqrt{nh_1}h_1^{\rho_1+1}$.

The default asymptotic exercise for this section has the sample size growing large but the number of cutoffs fixed. In the third case of heterogeneity assumptions, the treatment function is an unknown object of only finite dimension, and we do not need the number of cutoffs to grow to infinity to approximate the integral average treatment effect. Nevertheless, under conditions similar to Theorem 2, we also obtain asymptotic normality of $\widehat{\boldsymbol{\theta}}$ under the asymptotics with a large number of cutoffs.

Corollary 1. *Assume conditions in 1, 2, 3, 4, 5(a,b), 6(a,c,d,e) hold. Assume 7 holds and its vector valued function* $\boldsymbol{\mathcal{W}}(c,d)$ *is bounded in* $\mathcal{X}\times\mathcal{D}$. *Assume 6(f) holds for the* $q\times 1$ *vector valued weights*

$$w_{p,j}=\left(\mathbf{W}'\Omega\mathbf{W}\right)^{-1}\mathbf{W}'\Omega_{\bullet,(p,j)}$$

in the place of $\Delta_{p,j}$*, where* $\Omega_{\bullet,(p,j)}$ *is the column of* Ω *associated with cutoff* (p,j)*, and that* $\max_{p,j} \left\| \left(\mathbf{W}'\Omega\mathbf{W}\right)^{-1}\mathbf{W}'\Omega_{\bullet,(p,j)} \right\| = O(K^{-1})$*. As* $n \to \infty$*, assume* $K \to \infty$*, and* $h_1 \to 0$ *such that*

- $\frac{\sqrt{K}\log n}{\sqrt{nh_1}} \to 0$ *and* $Kh_1 = O(1)$
- $\sqrt{Knh_1}h_1^{\rho_1+1} \to C \in [0,\infty)$ *where* ρ_1 *is the order of the first stage LPR*

then

$$\sqrt{Knh_1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(C \cdot AB_\theta; AV_\theta\right)$$
$$\sqrt{Knh_1}\left(\widehat{\mu}(F) - \mu(F)\right) \xrightarrow{d} N\left(C\ \mathbf{Z}(F)\ AB_\theta; \mathbf{Z}(F)\ AV_\theta\ \mathbf{Z}(F)'\right)$$

where

$$AB_\theta = \lim_{K,n\to\infty} \sum_{p,j} w_{p,j}\mathcal{B}_{p,j}$$
$$AV_\theta = \lim_{K,n\to\infty} K \sum_{p,j} w_{p,j}w'_{p,j}\mathcal{V}_{p,j}$$

where $\mathcal{B}_{p,j}$ *and* $\mathcal{V}_{p,j}$ *are defined in equations 11 and 12.*

Both estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mu}(F)$ have a faster convergence rate under the large K asymptotics. Differently than Theorem 2, there is no lower bound requirement on the speed that K grows relatively to n which includes the fixed K case. The finite sample expressions for bias and variance that one obtains from Theorem 3 or Theorem 1 do not change. Consistent estimators for the asymptotic bias and variance terms are constructed using the estimators $\widehat{\mathcal{B}_{p,j}}$ and $\widehat{\mathcal{V}_{p,j}}$ proposed in the end of section 2. Sufficient moment and rate conditions give consistency of these estimators under the asymptotics with large number of cutoffs.

4.2 Fuzzy RDD

Another key advantage of the third case of heterogeneity assumptions is that a parametric functional form obtains identification in the fuzzy RDD case with multiple cutoffs. In the sharp RDD case, all individuals with the same forcing variable x and in the same subpopulation p receive the same treatment $D(x, p)$ (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.[9] 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.

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. If an individual is eligible to receive a treatment dose $d \in \mathcal{D}$, the ideal treatment dose this individual would receive is given by the random variable $\mathcal{U}_i(d)$ defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For example, if an individual i (or $\omega \in \Omega$) is eligible to a high school of quality $d \in \mathcal{D}$, then $\mathcal{U}_i(d) = \mathcal{U}_i(\omega, d) \in \mathcal{D}$ denotes the ideal quality of a high school this individual would like to attend. In practice, there is only a finite number of schools to go to. The quality of the high school an individual attends is the closest quality to her ideal quality given her eligibility.

$$\bar{\mathcal{U}}_i(d, p) = \min \left\{ \underset{u \in \{d_{0,p}, \ldots, d_{K(p),p}\}}{\operatorname{argmin}} |\mathcal{U}_i(d) - u| \right\} \tag{18}$$

[9] The RDD assignment is fuzzy for application-specific reasons. 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. We hold on to the simple example of the high school assignment problem in the main text for ease of exposition.

where the min{.} picks the smallest d in case of a tie, and $\bar{\mathcal{U}}_i(d,p)$ is a random variable defined on $(\Omega, \mathcal{A}, \mathbb{P})$ that depends on the schedule of cutoffs and treatment doses. We do not observe the potential treatment assignments $\bar{\mathcal{U}}_i(d,p)$ but just the actual treatment dose received.

$$D_i = \bar{\mathcal{U}}_i\left(D(X_i, P_i), P_i\right)$$

where $D(x,p)$ determines the treatment dose of eligibility (defined in eq. 1).

We build on classical definitions of compliance behaviors (e.g. Imbens and Rubin (1997)) and define three types of compliance groups. We use a simple example with 3 schools in one sub-population ($P_i = 1, K = 2$) to introduce the different compliance behaviors. The qualities of these schools are denoted as $d_0 = 0$, $d_1 = 1$, $d_2 = 2$. Table 1 lists all possible combinations of treatment eligibility and assignment.

Table 1: Different Compliance Behaviors

Eligibility			**Type**
0	**1**	**2**	
0	0	0	never-changers
1	1	1	
2	2	2	
0	0	2	ever-compliers
0	1	1	
0	1	2	
1	1	2	
0	0	1	ever-defiers
0	1	0	
0	2	0	
0	2	1	
0	2	2	
1	0	0	
1	0	1	

Eligibility			**Type**
0	**1**	**2**	
1	0	2	ever-defiers
1	1	0	
1	2	0	
1	2	1	
1	2	2	
2	0	0	
2	0	1	
2	0	2	
2	1	0	
2	1	1	
2	1	2	
2	2	0	
2	2	1	

Notes: All possible realizations of the random function $\bar{\mathcal{U}}_i(\cdot, 1)$ which denotes the treatment dose received given the eligibility for treatment doses 0, 1, and 2.

The three different behaviors 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 better treatment dose they become eligible for. 'Ever-defiers' change to a treatment dose different from one they are eligible for. In the case of two schools, our definition is equivalent to the classical definition of compliers and defiers.

In the high school assignment case, an example of a 'never-changer' is a student who strongly prefers the high school with the lowest admission cutoff and will attend that high school even if she is admitted to better schools. An example of a '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 can always pick 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.

The three compliance groups are measurable sets that partition the Ω space and are defined in terms of the ideal treatment function $\mathcal{U}_i$.[10]

$$\mathbf{G}_{nc} = \{\omega \in \Omega : \ \forall d \neq d' \in \mathcal{D}, \ \ \mathcal{U}_i(d) = \mathcal{U}_i(d')\} \tag{19}$$

$$\mathbf{G}_{ec} = \{\omega \in \Omega : \ \forall d \neq d' \in \mathcal{D}, \ \ \mathcal{U}_i(d) = \mathcal{U}_i(d') \text{ or } \mathcal{U}_i(d) < \mathcal{U}_i(d') = d'\} \bigcap \mathbf{G}_{nc}^c \tag{20}$$

$$\mathbf{G}_{ed} = \Omega \bigcap \mathbf{G}_{ec}^c \bigcap \mathbf{G}_{nc}^c \tag{21}$$

[10]We are implicitly assuming that the set of all possible 'ideal treatment' functions $\{\mathcal{U}_i(\omega, \cdot), \omega \in \Omega\}$ is such that the compliance sets are measurable. Moreover, the definitions of the three groups is based on the treatment doses being increasing in j for each sub-population, $d_{p,0} < d_{p,1} < \ldots < d_{p,K(p)}$. The definitions can be changed to accommodate decreasing or non-monotonic treatment doses. For example, if a school with a higher admission cutoff happens to have a lower quality, or in the class-size rule applications when the class size drops after each cutoff. Ever-compliers could then be defined as those who comply at least once to some eligibility change no matter what dose they receive prior to the change.

where A^c indicates the complement of set A.

Our definition of the compliance groups does not depend on the schedule of cutoffs and treatment doses, but only on $\mathcal{U}$, an intrinsic characteristic of each individual. It is plausible to assume that the treatment received never changes unless it changes to comply with the change in treatment eligibility. This minimal assumption rules out ever-defiers in the population. For never-changers, we can never identify treatment effects because they never undergo a change in treatment dose. Because of the multiplicity of treatments, ever-compliers can differ from each other when it comes to the number of 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.

Assumption 8 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 (2015). It also states that the functional form of treatment effects does not vary across different types of ever-compliers. We do not observe potential treatment assignments but only the treatments individuals actually receive. Therefore, we cannot distinguish one type of ever-complier from another.

Assumption 8. *Fix arbitrary $d \in \mathcal{D}$, and $p \in \mathcal{P}$, $x \in \mathcal{X}$.*

(i) There are no ever-defiers: $\mathbb{P}[\mathbf{G}_{ed}] = 0$

(ii) For $\forall G \in \mathcal{A}$ *such that* $G \subseteq \mathbf{G}_{ec} \bigcup \mathbf{G}_{nc}$, $\mathbb{E}[Y_i(d)|X_i = x, P_i = p, G]$ *and* $\mathbb{P}\left[G|X_i = x, P_i = p\right]$ *are continuous functions of* x.

(iii) $\beta_{ec}(\mathbf{c}, p) \equiv \mathbb{E}[Y_i(d') - Y_i(d)|X_i = x, P_i = p, G] \quad \forall G \in \mathcal{A} : G \subseteq \mathbf{G}_{ec}$

In the example of table 1, we have two cutoffs. After we rule out ever-defiers, we are left with 3 possible treatment effects on ever-compliers: $\beta_{ec}(c_1, d_0, d_1)$ at cutoff c_1; $\beta_{ec}(c_2, d_0, d_2)$

and $\beta_{ec}(c_2, d_1, d_2)$ at cutoff c_2. Comparing the average outcome of individuals around cutoff c_1 identifies the treatment effect $\beta_{ec}(c_1, d_0, d_1)$. Identification is not possible at cutoff c_2 because there are two treatment effects to be retrieved from one observed change in average outcome around such cutoff. It becomes necessary to impose some assumption on how the identified treatment effect at cutoff c_1 relates to the unidentified effects in cutoff c_2. The parametric functional form of assumption 7 is sufficient for identification.

Similarly to the sharp case, assumption 1 applied to β_{ec} allows one to pool observations from many sub-populations. Lemma 5 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 that cutoff. If the treatment effect function is linear in a vector of parameters, then the average change in outcomes at the cutoff is also linear in those parameters.

Lemma 5. *Under assumption 8,*

$$B_{p,j} = \sum_{\substack{l \in \mathcal{J}_p^0 \\ l<j}} \omega_{p,j,l} \beta_{ec}(c_{p,j}, d_{p,l}, d_{p,j}, p)$$

where

$$\omega_{p,j,l} = \lim_{e \downarrow 0} \{\mathbb{P}[D_i = d_{p,l} | X_i = c_{p,j} - e, P_i = p] - \mathbb{P}[D_i = d_{p,l} | X_i = c_{p,j} + e, P_i = p]\} \quad (22)$$

and $B_{p,j}$ is defined in equation 4. Moreover, suppose assumption 7 hold for β_{ec}. Define

$$\widetilde{W}_{p,j} = \sum_{l \in \mathcal{J}_p^0, l<j} \omega_{p,j,l} \left[\boldsymbol{\mathcal{W}}(c_{p,j}, d_{p,j}) - \boldsymbol{\mathcal{W}}(c_{p,j}, d_{p,l})\right] \quad (23)$$

for the vector basis function $\boldsymbol{\mathcal{W}}(c, d)$ of assumption 7; and define the $K \times q$ matrix $\widetilde{\mathbf{W}}$ by stacking $\widetilde{W}_{p,j}$, and $\mathbf{B}$ by stacking $B_{p,j}$. If $\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}}$ is invertible, then $\boldsymbol{\theta}_0^{ec}$ is identified and equal to $\left(\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}}\right)^{-1}\widetilde{\mathbf{W}}'\mathbf{B}$.

The ATE over ever-compliers is simply the expected value of β_{ec} under the counterfactual distribution F choose by researcher.

$$\begin{aligned}\mu^{ec}(F) &= \mathbb{E}_F\left[\beta_{ec}(\mathbf{c};\boldsymbol{\theta}_0^{ec})\right] \\ &= \mathbb{E}_F\left[\boldsymbol{\mathcal{W}}(c,d') - \boldsymbol{\mathcal{W}}(c,d)\right]'\boldsymbol{\theta_0}^{ec} \equiv \mathbf{Z}(F)\boldsymbol{\theta}_0^{ec}\end{aligned} \tag{24}$$

for the same $\mathbf{Z}(F)$ defined in equation 17.

Lemma 5 suggests a two-step estimation procedure for $\boldsymbol{\theta}_0^{ec}$ where first-step estimates of $B_{p,j}$ are regressed on estimates of $\widetilde{W}_{p,j}$. In the first-step, in addition to the estimates $\widehat{B}_{p,j}$ (equations 5, 6, and 7), we also obtain estimates $\widehat{\widetilde{W}}_{p,j}$ using LPRs of $\boldsymbol{\mathcal{W}}(c_{p,j},D_i)$ on X_i at each side of the cutoff $c_{p,j}$. For each $p \in \mathcal{P}$, $j \in \mathcal{J}_p$, and l-vector coordinate of $\widetilde{W}_{p,j}$, $l = 1,\ldots,q$, we compute

$$\widehat{\widetilde{W}}_{p,j}^{(l)} = \hat{a}_{p,j}^{(l)-} - \hat{a}_{p,j}^{(l)+} \tag{25}$$

$$\begin{aligned}(\hat{a}_{p,j}^{(l)+}, \hat{\mathbf{b}}_{p,j}^{(l)+}) = \underset{(a,\mathbf{b})}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{X_i - c_{p,j}}{h_1}\right) v_i^{p,j+} \\ \left[\boldsymbol{\mathcal{W}}(c_{p,j},D_i) - a - b_1(X_i - c_{p,j}) - \ldots - b_{\rho_1}(X_i - c_{p,j})^{\rho_1}\right]^2\end{aligned} \tag{26}$$

$$\begin{aligned}(\hat{a}_{p,j}^{(l)-}, \hat{\mathbf{b}}_{p,j}^{(l)-}) = \underset{(a,\mathbf{b})}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{X_i - c_{p,j}}{h_1}\right) v_i^{p,j-} \\ \left[\boldsymbol{\mathcal{W}}(c_{p,j},D_i) - a - b_1(X_i - c_{p,j}) - \ldots - b_{\rho_1}(X_i - c_{p,j})^{\rho_1}\right]^2\end{aligned} \tag{27}$$

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

weighted least squares problem:

$$\widehat{\boldsymbol{\theta}^{ec}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left(\widehat{\mathbf{B}} - \widehat{\widetilde{\mathbf{W}}}\boldsymbol{\theta} \right)' \Omega \left(\widehat{\mathbf{B}} - \widehat{\widetilde{\mathbf{W}}}\boldsymbol{\theta} \right) \tag{28}$$

Following equation 24, the estimator for $\mu^{ec}(F)$ 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 17.

We state sufficient conditions for asymptotic normality of these estimators in the default asymptotic exercise with a large number of observations but fixed number of cutoffs. Similar to the smoothness assumptions on the conditional mean of Y_i (assumption 4), we state smoothness assumptions on the conditional probabilities of treatment.

Assumption 9. *For arbitrary $p \in \mathcal{P}$, $d, d' \in \mathcal{D}$, and $G \in \mathcal{A}$ such that $G \subseteq \mathbf{G}_{ec} \bigcup \mathbf{G}_{nc}$,*

- $\mathbb{E}[Y_i(d)|X_i = x, P_i = p, G]$ *is $\rho_1 + 1$ times continuously differentiable w.r.t. x with $\rho_1 + 1$-th partial derivative $\nabla_x^{\rho_1+1}\mathbb{E}[Y_i(d)|X_i = x, P_i = p, G]$*
- $\mathbb{E}[Y_i(d)^2|X_i = x, P_i = p, G]$ *is a continuous function of x*
- *for $\widetilde{\mathbf{W}}$ defined in Lemma 5, $\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}}$ is invertible*
- $\mathbb{P}[G|X_i = x, P_i = p]$ *is $\rho_1 + 1$ times continuously differentiable w.r.t. x with $\rho_1 + 1$-th partial derivative $\nabla_x^{\rho_1+1}\mathbb{P}[G|X_i = x, P_i = p]$.*

Theorem 4. *Suppose assumption 7 hold for the treatment effect function on ever-compliers $\beta_{ec}(\mathbf{c})$. Suppose assumptions 2, 3, 8, 9 hold. As $n \to \infty$, $h_1 \to 0$, assume*

- $nh_1 \to \infty$
- $\sqrt{nh_1}h_1^{\rho_1+1} \to C \in [0, \infty)$

then

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

Moreover, the asymptotic 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* $\Omega = (C^2\mathcal{B}^{ec}\mathcal{B}^{ec\prime} + \mathcal{V}^{ec})^{-1}$. *Below, the definitions used:*

$$\mathcal{B}^{ec} = \left[\mathcal{B}^{ec}_{1,K(1)}, \ldots, \mathcal{B}^{ec}_{K(1),K(1)}, \mathcal{B}^{ec}_{1,K(2)}, \ldots, \mathcal{B}^{ec}_{K(P),K(P)}\right]'$$
$$\mathcal{V}^{ec} = diag\left\{\mathcal{V}^{ec}_{1,K(1)}, \ldots, \mathcal{V}^{ec}_{K(1),K(1)}, \mathcal{V}^{ec}_{1,K(2)}, \ldots, \mathcal{V}^{ec}_{K(P),K(P)}\right\}'$$
$$\mathcal{B}^{ec}_{p,j} = [1 \quad -\boldsymbol{\theta}_0']\boldsymbol{\Psi}_{p,j}$$
$$\mathcal{V}^{ec}_{p,j} = [1 \quad -\boldsymbol{\theta}_0']\boldsymbol{\Phi}_{p,j}[1 \quad \boldsymbol{\theta}_0']'$$
$$\boldsymbol{\Psi}_{p,j} = \frac{1}{(\rho_1+1)!}\left[\nabla_x^{\rho_1+1}\mathbf{m}(c^+_{p,j},p) - (-1)^{\rho_1+1}\nabla_x^{\rho_1+1}\mathbf{m}(c^-_{p,j},p)\right]e_1'\Gamma^{-1}\gamma^*$$
$$\boldsymbol{\Phi}_{p,j} = \frac{\boldsymbol{\zeta}^2(c^+_{p,j},p) + \boldsymbol{\zeta}^2(c^-_{p,j},p)}{f_{X|P}(c_{p,j},p)q_p}e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1$$
$$\mathbf{m}(x,p) = \mathbb{E}[\mathbf{Y}_i|X_i = x, P_i = p] - \sum_{j\in\mathcal{J}_p}\mathbb{I}\{c_{p,j}\le x\}\mathbf{J}_{p,j}$$
$$\mathbf{J}_{p,j} = \lim_{e\downarrow 0}\left\{\mathbb{E}[\mathbf{Y}_i|X_i = x+e, P_i = p] - \mathbb{E}[\mathbf{Y}_i|X_i = x-e, P_i = p]\right\}$$
$$\mathbf{Y}_i \equiv [Y_i \ \boldsymbol{\mathcal{W}}(c_{p,j}, D_i)]'$$
$$\boldsymbol{\zeta}^2(x,p) = \mathbb{E}\left[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|X_i = x, P_i = p\right]$$
$$\boldsymbol{\varepsilon}_i = \mathbf{Y}_i - \mathbb{E}[\mathbf{Y}_i|X_i, P_i]$$

$e_1,\ \Gamma,\ \gamma^*,\ \Delta$ *are defined in Theorem 1*

Estimates for the asymptotic bias and variance as well as the optimal weighting matrix are obtained in similar fashion as proposed in section 4.1, Theorem 3.

5 Application

In this section, we illustrate our methods using the data from PEU on the high school assignment in Romania.[11] PEU contributes to literature by providing rigorous evidence of the impacts of going to a better school on the academic performance of students and on the behavior of parents and teachers. To our knowledge, they were the first ones to apply RDD to a dataset with variation in cutoff-dose values much larger than a typical RDD application. As rich data become available, applications of RDD with many thresholds will become even more common reinforcing the already existing demand for our methods. Our purpose in this section is to show the empirical relevance of three of our main contributions:

(i) the interpretation of the average effect depends on the weighting scheme because local treatment effects are heterogeneous; we give one example of policy question where the average effect obtained by normalizing all cutoffs to zero does not provide the right answer;

(ii) many policy questions demand an ATE of a continuous counterfactual distribution; we give one example of such policy question and show that the estimated ATE provides a different answer than a simple average of local treatment effects;

(iii) a parametric functional form yields (i) an optimal weighting scheme that increases estimation precision; (ii) identification of treatment effects for the sub-group of 'ever-compliers' (fuzzy RDD case); we show that the optimal weighting scheme changes the inference conclusions, and that the effect on 'ever-compliers' is much larger than the

[11] The dataset is available online through the website of the *American Economic Review* where PEU was published.

Intent-to-Treat (ITT) effect which is obtained when the sharp RDD estimator is used in a fuzzy RDD application.

The administrative data from Romania covers 3 cohorts of 9 grade students for the years of 2001, 2002 and 2003. The size of the cohorts are 107812, 110912, and 115413, with a total of 334137 observations. We describe the essential elements of the high school assignment mechanism and refer the reader to PEU for the 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 eighth grade exam and the student's grade point average of grades 5-8. The Ministry of Education ranks students by their transition score and no other criteria. The mechanism assigns the student in the first place to her most preferred school, the student in the second place to her most preferred school, etc. Each school has a fixed number of vacancies, so the mechanism eventually reaches a student in the ranking whose most preferred school is full: it assigns this student to her most preferred school among those with vacancies. It is assumed that students always prefer the high schools in their home town to high schools in other towns which is reasonably the case for 13-14 year old kids living with their parents. Students cannot decline their assignment and have incentives to truthfully reveal their preference rankings.

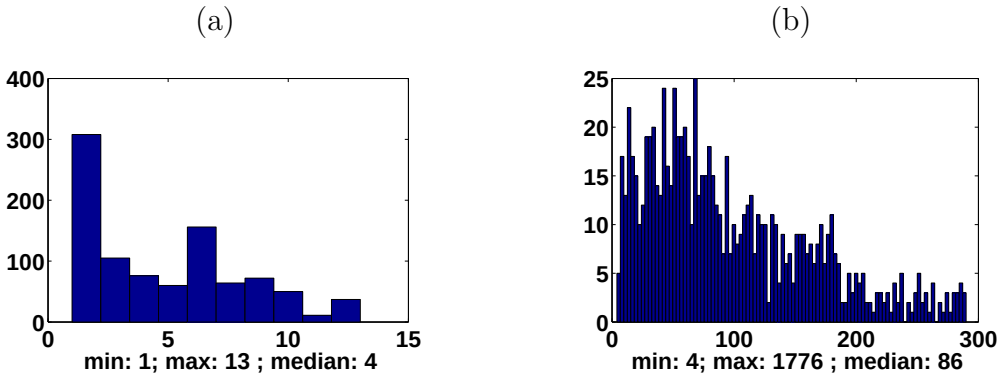
Another contribution of PEU is to analyze the data produced by such assignment mechanism using RDD methods. We observe the year, the town, the transition score X and the school each student is assigned to. The summary measure of school quality (treatment dose d) is the average-peer performance at each school. The treatment dose d is measured by the average transition score among those students that are assigned to that school. In translating this data to a RDD world, the cutoff for admission into a given school is computed by the minimum transition score among the students that are assigned to that school. For most students in the sample, we observe their score on the so called "baccalaureate exam" taken at the end of high school. The score on the baccalaureate exam is the outcome variable Y.

The ranking of preferences for high schools submitted by each student is not observed in the data which makes our RDD fuzzy. For example, a student could have a score greater than the cutoff for the school with highest d in her town but still be assigned to a different school because of her preferences. According to the cutoffs computed, 38% of the students in the sample are assigned to the high school with the highest d among those schools in their towns with admission cutoffs less or equal than their transition scores. If students had exactly the same rank of school preferences in each town, the RDD would be sharp and every student would attend the school with highest d among those with admission cutoffs less or equal than her score. Even in the fuzzy RDD case, the sharp treatment effect parameters of sections 2, 3, and 4.1 have the Intent-to-Treat (ITT) interpretation: they measure the average academic return of having access to a better school but not necessarily attending it. The fuzzy treatment effect parameters of section 4.2 measures the academic return of going to a better school averaged over the group of ever-compliers.

We compute the admission cutoffs and average peer-performance for each high school in each town-year. For a few town-years, the ordering of schools by admission cutoff does not correspond to the ordering by treatment dose d, a consequence of the fact that students' preference rankings over schools don't always coincide with the ranking of schools by average peer-performance d. We are interested in the effect of gaining access to a better school, so we only keep the cutoffs of those schools whose d is higher than the schools with smaller cutoffs. Also, we merge a few schools that happen to have the same admission cutoff in some town-years. These procedures lead to a monotonically increasing treatment schedule for every town-year $p \in \mathcal{P}$ $c_{p,j-1} < c_{p,j}$ and $d_{p,j-1} < d_{p,j}$ $\forall j \in \mathcal{J}_p$. Once we drop the observations with missing values for the baccalaureate exam, we are left with a total of 237062 students, 826 schools, 131 towns, and 939 cutoffs. Figure 2(a) illustrates the distribution of the number of cutoffs across town-years. 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.

In other words, each cutoff has a maximum estimation window around it such that windows do not overlap across cutoffs. For the majority of cutoffs in the sample, there are enough for observations for feasibility of first-step estimation (Figure 2(b)).

Figure 2: Number of Cutoffs and Observations per Cutoff



Notes: (a) Histogram of the number of cutoffs per town-year: in each town-year, the admission cutoff of a given school is the minimum transition score among the students that are assigned to that school. We only keep those cutoffs that grant access to a high school of higher quality.
(b) Histogram of the number of observations per cutoff: these are the observations available in the first-step estimation of $B_{p,j}$ for each cutoff (p, j). The neighborhoods around the cutoffs do not overlap and each individual observation in the sample is matched to only one cutoff.

For ease of exposition, we impose a restriction on the treatment effect function

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

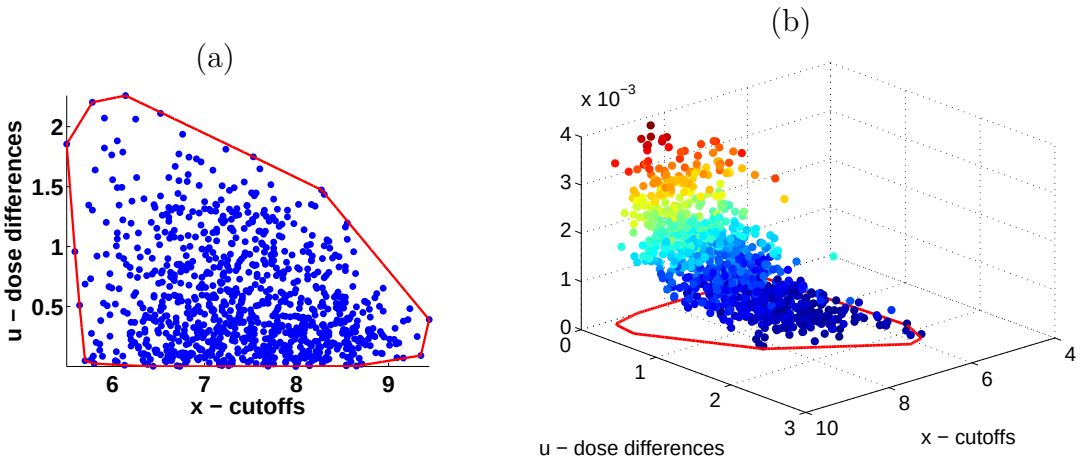
to reduce the dimension of its domain and make possible illustrations of functions of (x, u) in three dimensional graphs.[12] Figure 3(a) illustrates the variation in cutoff and dose-difference values (x, u) for the Romanian data. The convex hull of $\mathcal{C}_K = \{(x_{p,j}, u_{p,j})\}_{p \in \mathcal{P}, j \in \mathcal{J}_p}$ is our set $\mathcal{C}$ over which we compute average effects.

The common practice normalizes all cutoffs to zero and use the RDD estimator for one cutoff. Local effects are very likely to be heterogeneous, and the normalization procedure esti-

[12]This assumption restricts the returns of going to a better school to be linear in the average peer-performance d. The theory developed in this paper is general enough to deal with the three dimensional domain of β where returns to school quality do not have to be linear in d.

mates a weighted average of local treatment effects weighted by the relative density of individuals near each of the cutoffs (Cattaneo, Keele, Titiunik, and Vazquez-Bare (2015)). Although such implicit weighting scheme is often ignored in applied work, the interpretation such an average depends crucially on how local treatment effects are combined. Figure 3(b) shows the relative density of individuals around each cutoff, that is, $f_{X,P}(c_{p',j'}, p')/\sum_{p,j} f_{X,P}(c_{p,j}, p)$ for every (p', j') in the set $\mathcal{C}$ of figure 3(a). This is the implicit weighting scheme of the average effect obtained by the normalization procedure.

Figure 3: Cutoff-dose Values and Implicit Weighing of Normalization Procedure



Notes: (a) Scatter plot of cutoff and dose-difference values $\mathcal{C}_K = \{(x_{p,j}, u_{p,j})\}_{p\in\mathcal{P}, j\in\mathcal{J}_p}$ for the 939 cutoffs from the Romanian data. The line that envelops the scatter plot is the convex hull of the set of cutoff-dose values or the set $\mathcal{C}$ over which we compute ATEs.
(b) Three dimensional scatter plot of the implicit weighting scheme of the normalization procedure plotted over set $\mathcal{C}$. The Z-axis is the relative density of the forcing variable $f_{X,P}(c_{p',j'}, p')/\sum_{p,j} f_{X,P}(c_{p,j}, p)$ for every (p', j') in the set $\mathcal{C}$.

Assuming we are in the first case of heterogeneity assumptions, we compare the average estimate obtained by the normalization strategy with the ATE obtained by two different discrete counterfactual distributions (table 2). $F^{disc}_{(i)}$ corresponds to the implicit weighting scheme of the normalization strategy, that is, the relative density of the forcing variable at the existing cutoffs. This justifies the similar results of lines 1 and 2 of table 2. The average return of having access to a better high school for those students near the cutoffs is about

0.04 of a point in the baccalaureate exam grade (grades vary between 0 and 10). None of the estimates throughout this section are bias corrected because the bias terms are negligible. $F^{disc}_{(ii)}$ corresponds to a policy question described below.

Suppose that Romania wants to invest in elite high schools as part of a national science and innovation program. The new policy marginally increases the number of vacancies of high schools that currently admit the top 25% students, that is, those students with transition scores greater than 8.69. The goal is to grant high ability students access to better schools by marginally decreasing the admission cutoffs of elite high schools. Opponents to the schools' expansion argue that top quartile students who would be allowed into the most elite schools would not benefit sufficiently to justify the costs of modifying school buildings, transferring teachers, and the like. We are interested in the average effect among those students that are granted access to better high schools because of this policy. The distribution $F^{disc}_{(ii)}$ corresponds to the relative density of individuals that are local to the existing cutoffs of the selected high schools, and its estimate is shown in table 2, third line. The impact of this policy is statistically equal to zero which strongly suggests that local treatment effects are heterogeneous. High ability students won't benefit from having access to better schools, and conclusions based on the normalization procedure are misleading to answer such a policy question. The difference in the ATE due to different weighting scheme is statistically different than zero (table 2, 4th line).

Table 2: Heterogeneity Case I - Different Weighing Schemes

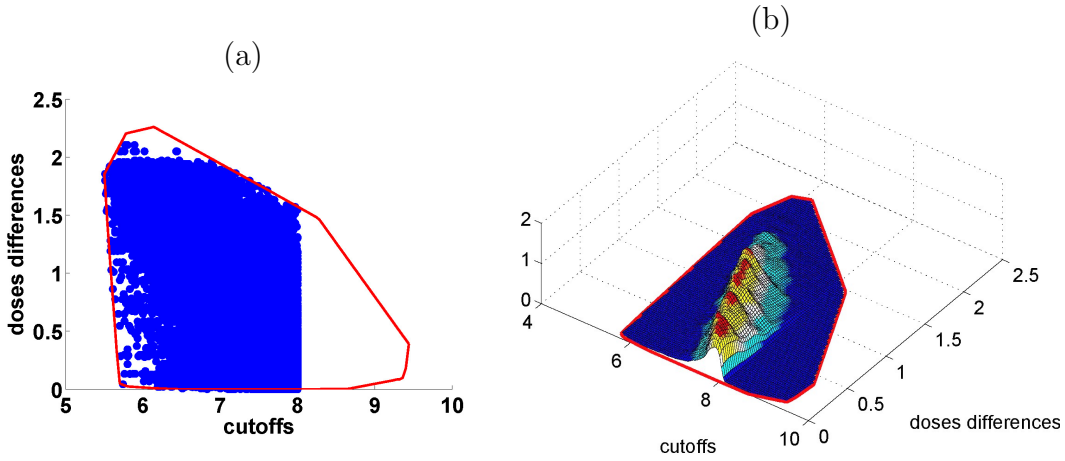
Method	Parameter	Estimate	S.E.
Normalization	$\mu(F^{disc}_{(i)})$	0.0381	0.0082***
Two-step	$\mu(F^{disc}_{(i)})$	0.0417	0.0113***
	$\mu(F^{disc}_{(ii)})$	-0.0243	0.0282
	$\mu(F^{disc}_{(i)}) - \mu(F^{disc}_{(ii)})$	0.0659	0.0276**

Notes: 'Normalization' pools data from all town-years where each individual is matched to his nearest admission cutoff and the cutoff value is subtracted from the individual transition scores X_i, so to have one cutoff at zero for everybody. Each individual observation is used only once in the estimation (no overlapping estimation windows). Normalization estimates are obtained by Local Linear Regression (LLR) with optimal IK bandwidth (Imbens and Kalyanaraman (2012)) and Edge kernel. Estimates for lines 2,3 and 4 are obtained according to the 2-step estimation procedure described in section 2. The first step uses LLR with IK bandwidth for almost all cutoffs, and the Nadaraya-Watson (i.e. $\rho_1 = 0$) for those few cutoffs that did not have enough observations to run a LLR. The weighting scheme (i) is $\omega_{p',j'} = \hat{f}_{X,P}(c_{p',j'}, p') / \sum_{p,j} \hat{f}_{X,P}(c_{p,j}, p)$ for every p', j', where $f_{X,P}(c_{p',j'}, p')$ is estimated using an uniform kernel with the Silverman's bandwidth. The weighting scheme (ii) is $\omega_{p',j'} = \hat{f}_{X,P}(c_{p',j'}, p') / \sum_{p,j \,:\, c_{p,j} \geq 8.96} \hat{f}_{X,P}(c_{p,j}, p)$ for p', j' such that $c_{p',j'} \geq 8.96$ and $\omega_{p',j'} = 0$ otherwise.

The inference method developed under heterogeneity case II allows for estimation of ATEs over continuous counterfactual distributions of the forcing variable and dose-changes. This permits inference over a much more general set of counterfactuals and not only those policies that target the individuals near the existing cutoffs. Suppose we are interested in using the Romanian data to predict how students would benefit from a new charter school that admits students from disadvantaged backgrounds. 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 less or equal than 8. Assume that, because the new charter school has more autonomy and better management than traditional public schools, it will be equivalent to a high school of average peer performance equal to 8 (even though its average student scores less than 8). Given these parameters, we compute the distribution of transition scores X, and dose changes $U = 8 - D$ of those individuals admitted into this charter school. Figure 4 illustrates the weighting density $\omega(x, u)$ implied by this policy counterfactual. Note that the

support of $\omega(x, u)$ involves not only individuals with transition scores equal to the observed cutoff dose-change values but also away from them (compare figures 3(a) and 4(a)).

Figure 4: Weighing Density of Charter School Example



Notes: (a) The contour line indicates the boundary of the set $\mathcal{C}$. The shaded region inside set $\mathcal{C}$ is made of the scatter plot of the transition scores and dose-change values of those individuals admitted into the charter school. The shaded area illustrates that the support of the weighting density $\omega(x, u)$ is more general than $\mathcal{C}_K$, the cutoff dose-change values observed in the data (figure 3(a)).
(b) The weighting density $\omega(x, u)$ implied by the charter school example.

We compute the estimate for $\mu(F^{cont})$ where F^{cont} is the cumulative distribution function associated with the density $\omega(x, u)$ of students admitted into the charter school (table 3). The estimate $\widehat{\mu}(F^{cont})$ is equal to 0.0547 of a point in the baccalaureate exam grade, and it is statistically different than zero. Estimation of $\mu(F^{cont})$ requires corrected weights as discussed in section 3. A natural question that arises is how well the ATE is approximated by an average of local effects. In other words, how well the parameter $\mu(F^{disc}_{(iii)})$ approximates the parameter $\mu(F^{cont})$, where $F^{disc}_{(iii)}$ is the discrete distribution with support $\mathcal{C}_K$ defined using the relative density weights from the charter school density $\omega(x, u)$. The estimate $\widehat{\mu}(F^{disc}_{(iii)})$ computed in this manner is approximately half of $\widehat{\mu}(F^{cont})$ and statistically insignificant. Also, under finite K asymptotics, the null hypothesis of equality between $\mu(F^{disc}_{(iii)})$ and $\mu(F^{disc}_{(iv)})$, where $F^{disc}_{(iv)}$ is defined using corrected relative density weights, is rejected at 5% significance (third line of table 3). The charter school policy question demands an ATE

computed over the entire distribution of students admitted. Using the distribution of students near the existing cutoffs in the Romanian data leads to misleading conclusions. The difference between and the ATE and the average of local effects arises from non-linearities in the treatment effect function $\beta(x,u)$ that are not captured in an average of local effects because it averages over only those individuals that are local to existing cutoff values.

Table 3: Heterogeneity Case II - Charter School Example

Parameter	Estimate	S.E.
$\mu(F^{cont})$	0.0547	0.0187***
$\mu(F^{disc}_{(iii)})$	0.0283	0.0176
$\mu(F^{disc}_{(iv)}) - \mu(F^{disc}_{(iii)})$	0.0264	0.0135**

Notes: The first step estimation uses LLR with IK bandwidth for almost all cutoffs, and the Nadaraya-Watson ($\rho_1 = 0$) for those few cutoffs that did not have enough observations to run a LLR. The second step estimation of $\mu(F^{disc}_{(iii)})$ averages first step estimates $\widehat{B}_{p,j}$ using the weighting scheme $\omega_{p',j'} = \omega(c_{p',j'}, p')/\sum_{p,j}\omega(c_{p,j},p)$, where $\omega(x,u)$ refers to the charter school density. The second step estimation of $\mu(F^{cont})$ uses a bivariate local cubic regression ($\rho_2 = 3$) to compute the corrected weights $\Delta_{p,j}$. Corrected weights depend on a choice of second step bandwidth h_2, and h_2 was chosen to minimize the MSE of $\widehat{\mu}(F^{cont})$. Using the charter school density $\omega(x,u)$, we define $\mu(F^{disc}_{(iv)})$ with corrected relative density weights, and $\mu(F^{disc}_{(iii)})$ with relative density weights.

We illustrate our estimation procedure for heterogeneity case III by specifying the following parametric functional form for the treatment effect function.

$$\beta(x,u) = \theta_1 u + \theta_2 xu + \theta_3 x^2 u$$

This functional form is chosen to be linear in u in order to be consistent with the restriction that $\beta(x,d,d') = \beta(x,d'-d) = \beta(x,u)$ that was imposed in the beginning of this section. The quadratic term in the transition score x allows for varying marginal effects of ability on the returns to school quality. In table 4, we report the estimates for the $\boldsymbol{\theta}$ and $\mu(F^{cont})$ for the charter school weighting density. We compare estimates for two choices of weighting matrix Ω: (i) cutoffs are equally weighted; (ii) cutoffs are weighted by the inverse of their first step variance $\mathcal{V}_{p,j}$ (optimal weighting that minimizes variance). The precision of

the parameter estimates is greatly improved when the optimal weighting is used, and all parameter estimates become statistically significant. Thetas that are different zero suggest heterogeneity of treatment effects across cutoffs. According to this parametric functional form, the ATE for the charter school example is positive, significant and higher than before.

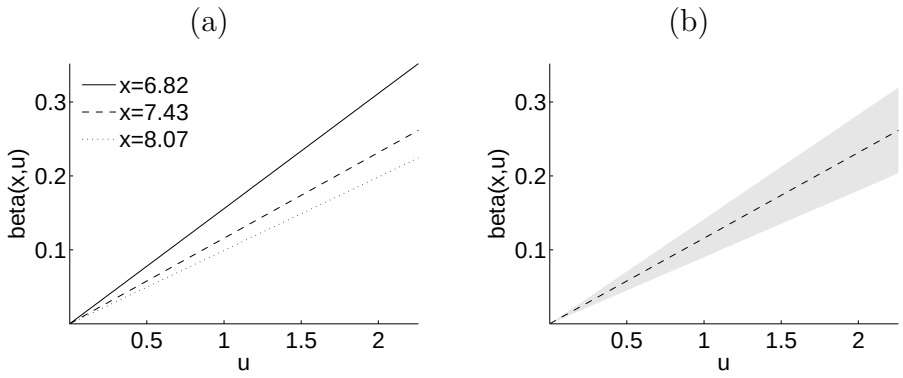
Table 4: Heterogeneity Case III - Sharp Case

	Equal Weighting		Optimal Weighting	
Parameter	Estimate	S.E.	Estimate	S.E.
θ_1	0.4802	1.2856	2.2085	0.5158***
θ_2	-0.1243	0.3509	-0.5171	0.1465***
θ_3	0.0100	0.0238	0.0317	0.0103***
$\mu(F^{cont})$ (charter school)	0.0790	0.0133***	0.1147	0.0084***

Notes: the first step estimates of $B_{p,j}$ were obtained by LLR with IK bandwidth for almost all cutoffs, and the Nadaraya-Watson ($\rho_1 = 0$) for those few cutoffs that did not have enough observations to run a LLR. Second step parametric estimates were obtained according to the procedure described in section 4.1 for two choices of weighting matrix Ω: (i) equal weighting, $\Omega = I_{K\times K}$ (identity matrix); (ii) optimal weighting, $\Omega = diag\{\hat{\mathcal{V}}_{p,j}\}_{p,j}$. The average $\mu(F^{cont})$ is the integral of the estimated parametric $\beta(x,u)$ weighted by the charter school weighting density $\omega(x,u)$.

Returns to better schooling are increasing in the change in school quality u, and the slope is larger for students with lower transition score. Figure 5(a) plots the treatment effect function $\beta(x,u)$ against u for the 25th, 50th, and 75th quantiles of the transition scores in set $\mathcal{C}$. Figure 5(b) repeats the plot of $\beta(x,u)$ for the median value of x and adds 95% pointwise confidence intervals.

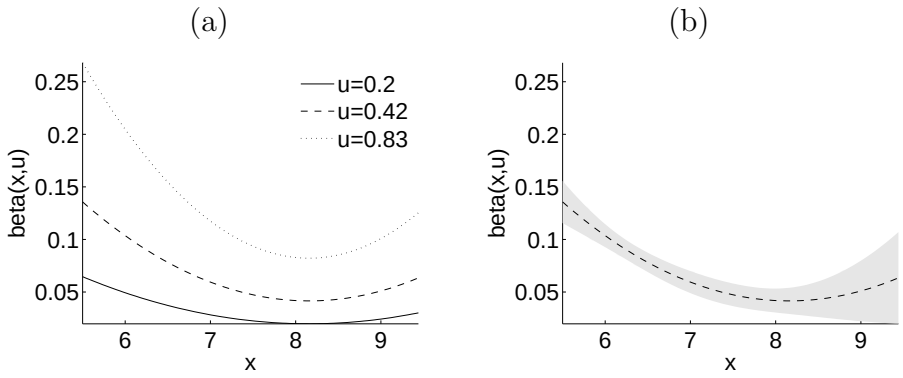
Figure 5: Returns to Better Peers and Change in Treatment Dose



Notes: (a) estimated parametric treatment effect function $\beta(x, u)$ plotted against changes in average peer performance u within set $\mathcal{C}$. Transition score x is fixed at three choices corresponding to the 25th, 50th and 75th percentiles of the distribution of cutoff values.
(b) For x fixed at the median cutoff value, $\beta(x, u)$ is plotted against u along with 95% confidence bands.

Returns to better schooling generally decrease as the transition score x increases, as shown in figure 6. Panel (a) plots the treatment effect function $\beta(x, u)$ against x for three fixed values of dose changes that are equal to the 25th, 50th and 75th quantiles of the dose-changes in the set $\mathcal{C}$. Panel (b) plots 95% confidence bands around the $\beta(x, u)$ for the median value of u.

Figure 6: Returns to Better Peers and Transition Score



Notes: (a) estimated parametric treatment effect function $\beta(x, u)$ plotted against transition score x within set $\mathcal{C}$. The change in average peer performance u is fixed at three choices corresponding to the 25th, 50th and 75th percentiles of the distribution of changes in average peer performance.
(b) For the median value of u, $\beta(x, u)$ is plotted against x along with 95% confidence bands.

Thus far the analysis has been restricted to ITT effects; that is, the average effect on the baccalaureate exam grade when students have access a high school with better peers. Using the parametric functional form specification from above, we turn to the fuzzy case where inference is conditioned on the subgroup of 'ever-compliers'. We compare estimates for two choices of weighting matrix Ω: (i) cutoffs are equally weighted; (ii) cutoffs are optimally weighted to minimize variance. Similar to the sharp case, the optimal choice of Ω reduces the standard-errors and changes the conclusion of the individual significance tests on the thetas (table 5). The marginal effects of x and u on treatment effects have a similar shape to the ones in the sharp case (figures 5 and 6). The ATE over students admitted to the charter school is much higher than the ITT ATE. Attending the better charter school has an impact on the baccalaureate grade of 0.07 point higher than the impact from only having access to the better charter school.

Table 5: Heterogeneity Case III - Fuzzy Case

Parameter	Equal Weighting		Optimal Weighting	
	Estimate	S.E.	Estimate	S.E.
θ_1^{ec}	-0.2458	1.6576	7.3906	0.7253***
θ_2^{ec}	0.1202	0.4362	-1.7812	0.1953***
θ_3^{ec}	-0.0093	0.0285	0.1080	0.0130***
$\mu^{ec}(F^{cont})$ (charter school)	0.1034	0.0176***	0.1868	0.0106***

Notes: the first step estimates of $B_{p,j}$ were obtained by LLR with IK bandwidth for almost all cutoffs, and the Nadaraya-Watson ($\rho_1 = 0$) for those few cutoffs that did not have enough observations to run a LLR. Second step parametric estimates were obtained according to the procedure described in section 4.2 for two choices of weighting matrix Ω: (i) equal weighting, $\Omega = I_{K\times K}$ (identity matrix); (ii) optimal weighting, $\Omega = diag\{\hat{\mathcal{V}}_{p,j}\}_{p,j}$. The average $\mu^{ec}(F^{cont})$ is the integral of the estimated parametric $\beta_{ec}(x,u)$ weighted by the charter school weighting density $\omega(x,u)$.

6 Conclusion

Regression discontinuity designs (RDD) have been used in a wide range of applications in Economics since the late 1990s. Identification and estimation results are well developed for the one cutoff case. More recently we see an increasing number of applications with one forcing variable and multiple cutoffs assigning individuals to heterogeneous treatments. There is a lack of theoretical studies investigating the conditions under which researchers can combine multiple local treatment effects to estimate an average treatment effect (ATE). A common practice is to normalize all cutoffs to zero and use the one cutoff estimator to obtain a summary effect. The average effect of the normalization strategy can lead to misleading conclusions if interest lies on average effects with different distributions of individuals including individuals away from existing cutoffs.

This paper proposes inference procedures for average effects in RDD with multiple thresholds. Our estimator is consistent and asymptotically normal for an average treatment effect over the entire support of variation in cutoffs and treatment doses. If treatment effects follow a non-parametric model, asymptotic results require both the number of observations and cutoffs to grow large. The rate of growth of the number of cutoffs relative to the number of observations determines the feasible set of bandwidth choices. The number of cutoffs cannot grow too fast to allow consistent estimation of local treatment effects uniformly across cutoffs. The number of cutoffs cannot grow too slowly to control the bias in the integral approximation of the ATE. The maximum rate of convergence of the estimator is root-n within the feasible set of bandwidth choices. If treatment effects follow a parametric model, then observations can be optimally combined for efficiency, and a parametric function form obtains identification in the fuzzy case.

We apply our methods to the data of PEU on high school assignment in Romania based on transition scores of students. We examine estimates for two types of average effects: (i) average of local treatment effects (ALTE); and (ii) average treatment effect (ATE). For the (ALTE), we compare the weighting scheme implicit to the normalization strategy (relative

density of individuals at the cutoff values) to a weighting scheme over the top 25% of individuals in the distribution of transition scores. Statistically different average effects illustrate the heterogeneity of local treatment effects and the inability of the normalization strategy to answer policy questions that lead to different weighting schemes. We estimate the ATE over a distribution of individuals admitted into a fictitious charter school. Results indicate the inability that an ALTE has to predict the effect of such policy.

We also illustrate estimates of a simple parametric specification that allows returns to better schooling to vary with the transition score and school quality. We find that the optimal weighting scheme that minimizes variance changes inference conclusions on individual parameters. Returns to better schooling are increasing in school quality but at a decreasing rate in transition score. The high school assignment in Romania translates into a fuzzy RDD, so we use the parametric specification to infer the effects on ever-compliers. We find that the average return of going to the charter school for ever-compliers is almost twice as big as the average return of merely having access to the charter school.

References

Andrews, D. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, ed. by R. Engle, and D. McFadden, vol. 4, chap. 37, pp. 2247–2294. North Holland.

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 (2013): "Wanna Get Away? RD Identification Away from the Cutoff," *IZA Discussion Paper 7429*.

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 (2015): "External Validity in Fuzzy Regression Discontinuity Designs," *Working Paper, Stanford University*.

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.

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 (2015): "Identification in Regression Discontinuity Designs with Multiple Cutoffs," *Working Paper, University of Michigan.*

De Giorgi, G., A. Drenik, and E. Seira (2015): "Moral Hazard in the Credit Card Market," *Work in progress, 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. (2014): "Jump or Kink? Identification of Binary Treatment Regression Discontinuity Design without the Discontinuity," *Working Paper, University of California, Irvine.*

——— (2015): "An Alternative Assumption to Identify LATE in Regression Discontinuity Designs," *Working Paper, UC Irvine.*

Dong, Y., and A. Lewbel (2015): "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," *Forthcoming, Review of Economics and Statistics.*

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), 1739–74.

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*.

Heckman, J., and E. Vytlacil (2007): "Econometric Evaluation Of Social Programs, Part I: Causal Models, Structural Models And Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6, chap. 70, pp. 4779–4874. North Holland.

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.

Lipman, Y., D. Cohen-Or, and D. Levin (2006): "Error Bounds and Optimal Neighborhoods for MLS Approximation," in *Eurographics Symposium on Geometry Processing*, ed. by K. Polthier, and A. Sheffer.

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 (2014): "Empirical Likelihood for Regression Discontinuity Design," *Journal of Econometrics, forthcoming*.

Pollard, D. (1984): *Convergence of Stochastic Processes*. Springer.

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, Department of Economics, University of Wisconsin at Madison*.

Rokkanen, M. (2014): "Exam Schools, Ability, and the Effects of Affirmative Action: Latent Factor Extrapolation in the Regression Discontinuity Design," *Job Market Paper, MIT*.

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

Silverman, B. W. (1986): *Density Estimation For Statistics And Data Analysis*, vol. 26. Chapman and Hall.

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

Van Der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer.

White, H., and K. Chalak (2013): "Identification and Identification Failure for Treatment Effects Using Structural Systems," *Econometric Reviews*, 32(3), 273–317.

The Appendix is available online at:

www.stanford.edu/∼bertanha/Bertanha_JMP_appendix.pdf