A Global Regression Discontinuity Design: Theory and Application to Grade Retention Policies*

Isaac M. Opper[†] Umut Özek[‡]

February 19, 2024

Abstract

We use a marginal treatment effect (MTE) representation of a fuzzy regression discontinuity setting to propose a novel estimator. The estimator can be thought of as extrapolating the traditional fuzzy regression discontinuity estimate or as an observational study that adjusts for endogenous selection into treatment using information at the discontinuity. We show in a frequentest framework that it is consistent under weaker assumptions than existing approaches and then discuss conditions in a Bayesian framework under which it can be considered the posterior mean given the observed conditional moments. We then use this approach to examine the effects of early grade retention. We show that the benefits of early grade retention policies are larger for students with lower baseline achievement and smaller for low-performing students who are exempt from retention. These findings imply that (1) the benefits of early grade retention policies are larger than have been estimated using traditional fuzzy regression discontinuity designs but that (2) retaining additional students would have a limited effect on student outcomes.

[*] We thank Christine Mulhern for the incredibly helpful conversations about the method. We also thank conference participants at 2023 AEFP for their thoughtful comments and the anonymous school district for providing the data used in the analysis. An R package which implements the model can be found at: https://github.com/isaacopper/GlobalRDD. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D00008. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

[†] RAND Corporation. Email: iopper@rand.org.

[‡] RAND Corporation. Email: uozek@rand.org.

I Introduction

Regression discontinuity (RD) designs have become increasingly popular in empirical research over the past three decades (Cook (2008), Abadie and Cattaneo (2018), Cattaneo and Titiunik (2022)). This framework leverages plausibly exogenous discontinuities in treatment likelihood at predetermined cutoffs to identify the causal effect of the treatment (Imbens and Lemieux (2008), Hahn et al. (2001)). When the discontinuity in treatment likelihood is fuzzy – i.e., some individuals on the treatment side of the cutoff do not receive treatment while some individuals on the other side receive treatment – a common approach is to use an instrumental variables (IV) design where being on the treatment side of the cutoff is used as an instrument for receiving treatment. While such fuzzy designs generally provide compelling evidence of the treatment effect, the IV estimator yields an average treatment effect that is local (i.e., the LATE) in two ways.

First, the estimated effects only apply to individuals for whom being on the treatment side of the cutoff determines treatment status (Bertanha and Imbens (2020)). Second, similar to sharp RD designs with perfect compliance, it is hard to generalize these estimates to individuals identified for treatment who are away from the cutoff (Angrist and Rokkanen (2015), Cattaneo et al. (2021), Dong and Lewbel (2015)). Yet, understanding treatment effects beyond compliers at the cutoff is important from a public policy perspective. For example, moving beyond the LATE is necessary if one wants to (1) assess how increasing compliance among those identified for treatment may influence the effectiveness of the policy; (2) understand whether exemptions often incorporated into public policies indeed identify individuals less likely to benefit from treatment; or (3) examine the effect of the treatment on individuals away from the cutoff who typically have higher needs (e.g., educational), which is essential to assess the overall benefits of the policy.

Given these limitations, several recent studies have proposed methods to improve the external validity of the RD estimator. These approaches generally rely on strong assumptions such as the fact that individuals do not endogenously select into treatment (e.g., Angrist and Rokkanen (2015) and Bertanha and Imbens (2020)) or on the presence of multiple discontinuities (e.g., Cattaneo et al. (2021)). In practice, therefore, researchers often face an unenviable choice of whether to be satisfied with an (often noisy) estimate of the local treatment effect or the make strong (and often unbelievable) assumption that individuals do not endogenously select into the treatment.

This paper addresses the locality issue by introducing a new estimator for use in a fuzzy regression discontinuity setting that generates global treatment effect estimates. The estimator jointly models the two potential outcomes and selection into treatment. We then greatly restrict the set of potential functions to ensure that our estimator converges to a unique marginal treatment effect (MTE) function. Although identification is obtained by restricting the set of potential functions, we show that the estimator can be easily interpreted and motivated even in cases where the true MTE function does not lie in this restricted set. For example, our method corresponds to a linear combination of the estimate generated using a traditional fuzzy regression discontinuity design and the estimate using a traditional observational study, with the weights depending on how biased the traditional observational study appears to be based on behavior at the discontinuity.[1] In a similar fashion, the estimator can be thought of as an extrapolation of the traditional fuzzy regression discontinuity estimate, where the initial extrapolation to the non-compliers at the threshold is done using existing approaches usually employed in the RCT setting (e.g., Brinch et al. (2017); Kowalski (2023)) and extrapolation away from the threshold is done using the assumption that the amount of endogenous selection stays constant.

We next turn to the formal motivation of the estimator. We first show the conditions under which it converges to the true marginal treatment effect function are weaker than existing approaches that aim to extrapolate fuzzy RDD estimates. In particular, rather than assuming that there is *no* endogenous selection of individuals into treatment, we allow for individuals to sort into treatment based on information we do not observe and instead require that the way they do so does not vary away from the discontinuity and can be modeled using existing approaches to model imperfect compliance in the context of an RCT (e.g., Brinch et al. (2017); Kowalski (2023)).

Depending on the circumstances, however, it is plausible that researchers still may not feel comfortable making these assumptions. We therefore develop a Bayesian model, in which the true conditional moments are themselves distributed according

[1] Similarly the estimator can also be interpreted as starting with an observational study and then adjusting for bias using information at the discontinuity in a similar fashion as Bertanha and Imbens (2020).

to some prior distribution. We then discuss the conditions under which the proposed estimator can be thought of as the researchers "best guess" of the MTE function given the observed data, i.e., the mean posterior. For example, we show that unless the researcher has strong priors in how the MTE deviate from being linear functions of the variables, the proposed estimator gives the approximate mean posterior if one is only interested in extrapolating to non-compliers and to values of the running variable near the discontinuity.

We then use this estimator to examine the broader effects of early grade retention policies in the United States. This application is important for two reasons. First, this exercise has important implications for education policy in the United States: as of 2020, about half of all states and the District of Columbia require or encourage school districts to retain third-grade students who lag behind based on their third-grade reading scores. There is growing literature examining the effects of these policies using RD designs[2], yet we know very little about their effects on students away from the retention cutoff who have lower initial achievement.

Second, many early grade retention policies include "exemptions" to test score thresholds, such as for students who have disabilities, who are recent English learners, or whose proficiency can be demonstrated with a teacher's portfolio. As such, nearly all existing RD studies on early grade retention rely on a fuzzy RD design to identify the effect of retention on student outcomes. Yet we do not know if these exemptions indeed identify students who are less likely to benefit from retention. We examine these research questions using student-level administrative data from Florida, which requires third graders to score at or above Level 2 (out of 5 achievement levels) on the statewide reading test to be promoted to fourth grade.

Our findings suggest that the benefits of retention (1) are larger for students with lower baseline reading achievement and (2) are indeed smaller for students exempt from retention. Together, these results imply that the average treatment effect on the treated (ATT) is much larger than the predicted effects that would come from removing the exemptions or increasing the passing threshold, i.e., the average treatment effect on the control (ATC). For example, we find that, as currently implemented, retaining students increases their sixth grade reading scores by 0.69σ, but further in-

[2] For example, see Greene and Winters (2007), Winters and Greene (2012), Özek (2015), Schwerdt et al. (2017), Figlio and Özek (2020) in Florida; Hwang and Koedel (2022) in Indiana; and Mumma and Winters (2023) in Mississippi.

creasing the threshold by 50 points (0.8σ, roughly equivalent to moving the threshold from Level 2 to slightly above Level 3 on the third-grade reading test) and removing exemptions would have no impact on the sixth grade reading scores of the newly retained students. These findings also imply that existing studies on early grade retention policies that rely on traditional fuzzy RD designs significantly underestimate the benefits of retention. In particular, we show that the ATT estimates are roughly 20 percent larger than the LATE estimates of the retention effects on reading scores in grades 4 through 8.

II Model Assumptions and Estimation Approach

II.A Underlying Model and Assumptions

We use as our base model one of the canonical models used to consider the effect of a binary treatment on a single outcome, the model that forms the basis for marginal treatment effect (MTE) estimation (e.g., Heckman (2010); Heckman and Vytlacil (2007a,b); Brinch et al. (2017); Mogstad et al. (2018); Kline and Walters (2019)). Specifically, we assume that each individual is defined by four variables: their outcome if they are not treated, the effect that the treatment has on their outcome, their implied cost of enrolling in the treatment, and their value of the running variable; we denote these as μ_i, τ_i, η_i, and Z_i, respectively. In other words, we use μ_i to denote individual i's outcome in the absence of treatment and τ_i to denote the causal effect of the treatment on individual i's outcome; clearly $\mu_i + \tau_i$ is then their outcome if they are treated. In Section IV we discuss how to add additional covariates X_i to the analysis, but for expositional simplicity omit these for now.

Letting T_i be a dummy variable denoting whether someone is in the treatment or control group, the observed outcome can be written as: $Y_i = \mu_i + \tau_i T_i$. As is common in the MTE literature, we further assume that treatment is determined according the following choice equation: $T_i = \mathbf{1}(\nu(Z_i) \geq \eta_i)$ for some function of the running variable $\nu(Z_i)$. As a researcher, we observe Y_i, T_i, and Z_i, but do not observe the latent variables μ_i, τ_i, and η_i.

We then define following two conditional moments:

$$\mu^*(\eta, Z) = \mathbb{E}[\mu_i|\eta_i = \eta, Z_i = Z] \qquad \text{and} \qquad \tau^*(\eta, Z) = \mathbb{E}[\tau_i|\eta_i = \eta, Z_i = Z] \quad (1)$$

The function $\tau^*(\eta, Z)$, in particular, corresponds to the marginal treatment effect (MTE) function as defined in Heckman and Vytlacil (1999, 2005) and is generally the object of interest itself or, more commonly, the objects of interest can be derived from it. For example, full knowledge of the function $\tau^*(\eta, Z)$ would allow one to calculate the overall average treatment effect (ATE), the average treatment effect on the treated (ATT), and the average treatment effect on the compliers (LATE), and other estimands of interest. We use the star notation, i.e., denoting the functions as μ^* and τ^*, to distinguish the true conditional moment functions from generic potential conditional moment functions μ and τ.

While the conditional moment functions in Equation (1) correspond most closely with the objects of interest, they are a bit removed from what is observed in the data. We therefore also define two additional conditional moments, which are more closely related to what we observe. These moments are defined as follows:

$$y_0^*(\eta, Z) = \mathbb{E}[\mu_i|\eta_i > \eta, Z_i = Z] \qquad \text{and} \qquad y_1^*(\eta, Z) = \mathbb{E}[\mu_i + \tau_i|\eta_i \leq \eta, Z_i = Z] \quad (2)$$

These moments are redundant with the ones defined in Equation (1) as one can transform y_0 and y_1 to τ via the linear transformation:[3]

$$T(y_0, y_1) = y_1 - y_0 + \eta\frac{\partial y_1}{\partial \eta} + (1 - \eta)\frac{\partial y_0}{\partial \eta} \quad (3)$$

In the definitions above, we implicitly assume that the conditional first moments exist. We make this assumption explicit below, along with the other assumptions we use to capture the fuzzy regression discontinuity design.

Assumption 1. $\mathbb{E}[\mu_i|\eta_i = \eta, Z_i = Z] < \infty$ *and* $\mathbb{E}[\tau_i|\eta_i = \eta, Z_i = Z] < \infty$ *for all* $\eta_i \in (0, 1)$ *and* $Z_i \in \mathbf{Z} \equiv (\underline{Z}, \overline{Z})$.

Assumption 2. η_i *is continuously distributed conditional on* Z_i.

Assumption 3. *Both* $\mu^*(\eta, Z)$ *and* $\tau^*(\eta, Z)$ *are continuous functions of* (η, Z).

[3] Another way to write this transformation is that $T(y_0, y_1) = \frac{\partial}{\partial \eta}\Big(\eta y_1\Big) + \frac{\partial}{\partial \eta}\Big((1-\eta)y_0\Big)$. Similarly, there is way to transform y_0 and y_1 to μ, but this linear transformation is less important since researchers are generally interested in estimating the treatment effects.

Assumption 4. $\nu(Z) \in (0,1)$ *for all* Z *and is a continuous function at every point except for a single* Z^**, where:*

$$\lim_{Z \uparrow Z^*} \nu(Z) \equiv p_l < p_h \equiv \lim_{Z \downarrow Z^*} \nu(Z)$$

Assumption 5. *The researcher observes the conditional moments:* $\mathbb{E}[T_i|Z_i = Z]$, $\mathbb{E}[Y_i|T_i = 1, Z_i = Z]$, $\mathbb{E}[Y_i|T_i = 0, Z_i = Z]$ *at every point* $Z \neq Z^*$.

Assumptions 1 and 2 are relatively benign assumptions. The first makes explicit that our definitions of the conditional moments are valid and the second is a common assumption in the MTE literature and allows us to normalize η_i to be distributed uniformly from $(0,1)$. With this standard normalization, the cutoff value $\nu(Z_i)$ is equal to $Pr(T_i|Z_i)$, i.e. to the propensity score.

The next two assumptions are reformulations of the standard assumptions required for RD designs. Assumption 3 corresponds to the traditional assumption that the potential outcome functions are continuous around the discontinuity, although we extend the assumption to be that the functions are continuous everywhere and that they are continuous functions of both η_i and Z_i. Assumption 4 captures the fact that it is a fuzzy RD context, in that there is a single point Z^* at which the probability of treatment jumps discontinuously and that for every value of the running variable there are both treated and untreated individuals. We assume that there is a single discontinuity and that the probability increases as one moves across the threshold from left to right. The latter assumption is without loss of generality. The assumption that there is a single discontinuity is more consequential; however, one advantage of the method is that it can be naturally extended to cases in which there are multiple discontinuities. For ease of exposition, we focus on the case with a single discontinuity and discuss in Section B how the method can be adjusted when there are multiple discontinuities.

Finally, the last assumption states that the researcher observes the true conditional moments. While apparently quite a strong assumption, Assumption 5 is meant to clarify the main ideas by allowing us to focus on questions of identification in the main discussion. We then return to questions of estimation in Section IV, where we show that under a common set of assumptions we can (roughly speaking) invoke the law of large numbers to show that replacing the true expectations with estimates of these moments does not affect the main results, as long as we take an asymptotic perspective.

II.B Proposed Estimator

The key challenge is that – even when we ignore estimation error – we only observe the functions $y_k^*(\eta, Z)$ at a select number of points. To highlight this, consider any point $Z \neq Z^*$. As mentioned, we assume the researcher observes $\mathbb{E}[Y_i|T_i = 1, Z_i = Z]$ and $\mathbb{E}[Y_i|T_i = 0, Z_i = Z]$. From the definition of which potential outcome is realized and the specification that individuals select into treatment if (and only if) $\eta_i \leq \nu(Z_i)$, we can re-write the observed moment $\mathbb{E}[Y_i|T_i = 0, Z_i = Z]$ as:

$$\begin{aligned} \mathbb{E}[Y_i|T_i = 0, Z_i = Z] &= \mathbb{E}[\mu_i|T_i = 0, Z_i = Z] && (4) \\ &= \mathbb{E}[\mu_i|\eta_i > \nu(Z), Z_i = Z] && (5) \\ &= y_0^*(\nu(Z), Z) && (6) \end{aligned}$$

We can similarly re-write the observed moment $\mathbb{E}[Y_i|T_i = 1, Z_i = Z]$ as $y_1^*(\nu(Z), Z)$. Again focusing on a single Z, this formulation makes it clear how big the challenge is: we need to estimate the functions $y_0^*(\eta, Z)$ and $y_1^*(\eta, Z)$ when observing only a single point for each $y_0^*(\nu(Z), Z)$ and $y_1^*(\nu(Z), Z)$.[4] There is one exception to the above analysis: the point Z^* where $\nu(Z)$ jumps discontinuously. Instead of observing a single value of y_0^* and y_1^* at this point, we observe two. Specifically, we observe both:[5]

$$\lim_{Z \uparrow Z^*} y_0^*(\nu(Z), Z) \equiv y_0^*(p_l, Z^*) \qquad \text{and} \qquad \lim_{Z \downarrow Z^*} y_0^*(\nu(Z), Z) \equiv y_0^*(p_h, Z^*)$$

values of y_0^* at Z^*; similarly, we observe both $y_1^*(p_l, Z^*)$ and $y_1^*(p_h, Z^*)$.

Still, only observing two points at Z^* – and a single point everywhere else – means that we have no hope of non-parametrically identifying the functions y_0^* and y_1^*. An important implication of this is that the only way to identify the functions is via some functional form restriction on y_0^* and y_1^*. Our approach is to propose a particular restriction here – discussed below – and then spend the rest of the paper

[4] Of course, we observe these moments at multiple values of Z; however, that both provides an additional datapoint and also provides a different value of Z for which we need to project $y_0^*(\eta, Z)$ and $y_1^*(\eta, Z)$ and so it does not provide much help.

[5] Note that these limits are well-defined based Assumptions 1-4.

discussing ways to interpret and motivate the resulting estimand.

Our restriction will be that both $\hat{y}_0$ and $\hat{y}_1$ are additively separable and linear in η. To formally define the estimator, let $\mathcal{Y}^{GRDD}$ be the set of possible estimated functions $\hat{y} = (\hat{y}_0, \hat{y}_1)$. Then our restriction can be written as: $\mathcal{Y}^{GRDD} = \mathcal{Y}_0^{GRDD} \times \mathcal{Y}_1^{GRDD}$ where:

$$\mathcal{Y}_k^{GRDD} = \{y_k | y_k(\eta, Z) = \beta_k \eta + \gamma_k(Z) \text{ where } \beta_k \in \mathbb{R} \text{ and } \gamma_k(Z) \text{ is a continuous function.}\}$$

for $k \in \{0, 1\}$.

In other words, the restriction is that the conditional moments are additively separable and linear in η. One note about the terminology we use: we write that this is a restriction in the possible form that $\hat{y}_0$ and $\hat{y}_1$ take, rather than as an assumption. The reason is that some of the ways in which we motivate the estimator in Section III do not rely on the fact that the true functions take that form nor does the interpretation of the resulting estimates that we discuss in the next section.

With this definition of $\mathcal{Y}^{GRDD}$ we can now formally define our estimator, which we refer to as the "Global Regression Discontinuity Design" (Global RDD) and denote as τ^*_{GRDD}. The definition is as follows:

Definition 1. *Define τ^*_{GRDD} as $\tau^*_{GRDD} \equiv T(\hat{y}_0, \hat{y}_1)$ where $T(y_0, y_1)$ is defined as in Equation (3) and $\hat{y}_0$ and $\hat{y}_1$ are defined such that $(\hat{y}_0, \hat{y}_1) \in \mathcal{Y}^{GRDD}$ and $\hat{y}_k(\nu(Z), Z) = \mathbb{E}[Y_i | T_i = k, Z_i = Z]$ for all $Z \neq Z^*$ and $k \in \{0, 1\}$.*

There are two important points about the above definition. First, the estimator defined above is feasible, in that it does not rely on any data other than that which is assumed to be observed by the researcher under Assumption 5. Second, τ^*_{GRDD} is well-defined, in that under Assumptions 1-4 there is guaranteed to be a single value of τ^*_{GRDD} that meets that above definition. We state this as a proposition below:

Proposition 1. *Under assumptions 1-5, the Global RDD as defined in Definition can be implemented using the observed conditional moments and is well-defined.*

While we leave the formal proof to Appendix A, it is worth highlighting that the discontinuity at Z^* is precisely what ensures τ^*_{GRDD} to be well-defined; without the discontinuity, we would only observe a single point at every value of Z so even under the assumption that $y_k(\eta, Z) = \beta_k \eta + \gamma_k(Z)$ it would be impossible to pin down both β_k and the function $\gamma_k(Z)$ based on the observed data. We discuss this intuition in more detail and provide some alternative ways to view the estimator in the next subsection.

One final note about notation: although one advantage of the Global RDD is that it allows one to generalize both to the population of non-compilers and away from the discontinuity, it is often the case that researchers are particularly interested in understanding how the treatment varies with the running variable. We will use $\tau^*_{GRDD}(Z)$ as shorthand for the conditional average treatment effect (CATE), i.e., $\tau^*_{GRDD}(Z) \equiv \int_0^1 \tau^*_{GRDD}(Z, \eta)d\eta$.

II.C Discussion of the Estimator

Identification Intuition: To convey the intuition of how the Global RDD mechanically transforms the observed moments into the resulting estimates, we will consider a simplified example in which we are only concerned with the function $y_1(\eta, Z)$. The analysis for the function $y_0(\eta, Z)$ is identical and, as mentioned above, together the functions $y_1(\eta, Z)$ and $y_0(\eta, Z)$ pin down the MTE function $\tau(\eta, Z)$.

We illustrate the intuition using a stylized example in Figure 1. In Figure 1a, we start by illustrating the relationship between the running variable – shown on the x-axis – and the probability of treatment – shown on the y-axis. Note that we observe this function, i.e., $\mathbb{E}[T_i|Z_i = Z]$, in the data; in our notation, the line shown in Figure 1a corresponds to the function $\nu(Z)$. We also indicate ten points on the function with dots, which we use in the other figures, and label six of them.

In the next two panels, we turn our attention to the observed conditional means, i.e., to $\mathbb{E}[Y_i|Z_i = Z, T_i = 1]$. We plot these observed moments for the ten points we highlighted in the previous panel in Figure 1b, labelling the same six points as in Figure 1a, with the running variable (i.e, Z) on the x-axis. We do not draw a line through each of these points to emphasize that – unlike in Figure 1a – we are not directly concerned with how function $\mathbb{E}[Y_i|Z_i = Z, T_i = 1]$ varies as a function of Z. Instead, we are concerned with the question of how $y_1(\eta, Z)$ varies as a function of both η and Z. We could therefore similarly plot the observed points with the value of η, rather than Z on the x-axis. We do so in Figure 1c, again highlighting and labelling the same points as before.

It is this formulation that best highlights how the Global RDD transforms the observed moments into the resulting estimate of $\hat{y}_1(\eta, Z)$. To start, we will only concern ourselves with estimating the function $\hat{y}_1(\eta, 0)$. As discussed above, at this value of Z we observe two separate points: $\hat{y}_1(p_l, 0)$, which is identified as point C in Figure 1b, and $\hat{y}_1(p_h, 0)$, which is identified as point D in the Figure 1b. Without any restrictions, there are clearly many functions $\hat{y}_1(\eta, 0)$ that would go through both point C and D; if we restrict ourselves to linear functions, however, the two points completely determine the function $\hat{y}_1(\eta, 0)$.[6] This is shown in Figure 1d.

Of course, we also need to determine the functions $\hat{y}_1(\eta, -1)$, $\hat{y}_1(\eta, -0.9)$, ..., $\hat{y}_1(\eta, 0.8)$, $\hat{y}_1(\eta, 0.0)$, $\hat{y}_1(\eta, 1)$. If we restrict this set of functions to both all be linear functions of η as well as all have the same slope, i.e., for $\hat{y}_1(\eta, Z) = \beta\eta + \gamma(Z)$, then – after pinning down the slope using behavior at the discontinuity – we can adjust $\gamma(Z)$ such that $\hat{y}_1(\eta, Z)$ goes through every point. Again, we can see this by the functions $y_1(\eta, 1)$, $y_1(\eta, 0)$, and $y_1(\eta, -1)$ all consisting of parallel lines in Figure 1d.

As is clear, the only way that the Global RDD is able to transform the observed moments into estimates of $\tau^*(\eta, Z)$ is by greatly restricting the set of potential functions $y_1(\eta, Z)$. We want to emphasize, however, that as we discuss in Section III there are ways to motivate the estimator even if the true functions do not satisfy this restriction. First, however, we provide some more intuition about how the Global RDD mechanically transforms the observed moments into the resulting estimates by comparing it to alternative approaches.

Relationship to Alternative Approaches: For another perspective, we will consider two natural alternatives: (1) an observational study in which one simply compares the the treatment average to control average at every point $Z \in \mathbf{Z}$ and (2) a traditional fuzzy regression discontinuity design. Formally, we get that:

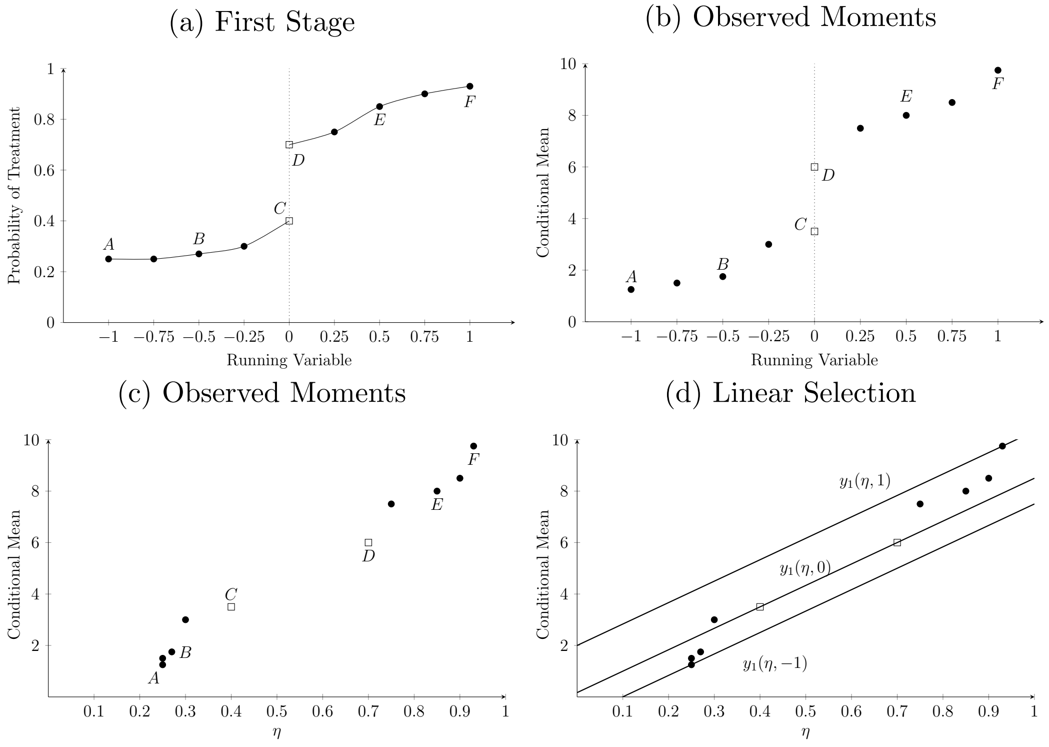
$$\tau^*_{obs}(Z) = \mathbb{E}[Y_i|T_i = 1, Z_i = Z] - \mathbb{E}[Y_i|T_i = 0, Z_i = Z] \tag{7}$$

$$\tau^*_{RDD} = \frac{1}{p_h - p_l}\left(\lim_{Z \downarrow Z^*} \mathbb{E}[Y_i|Z_i = Z] - \lim_{Z \uparrow Z^*} \mathbb{E}[Y_i|Z_i = Z]\right) \tag{8}$$

To show how the Global RDD transforms the observed moments into treatment effect estimates, we can then use the same stylized example as shown in Figure 1. In Figure 2a, for example, we plot directly the conditional moments as a function of the running variable; i.e., we plot $\mathbb{E}[Y_i|Z_i = Z, T_i = 1]$ and $\mathbb{E}[Y_i|Z_i = Z, T_i = 0]$ as a function of Z. In a traditional observational study, we then estimate the treatment

[6] This follows the discussion in Brinch et al. (2017).

Figure 1: Identification Intuition


(a) First Stage

(b) Observed Moments

(c) Observed Moments

(d) Linear Selection

Note: This figure illustrates the intuition of how the Global RDD transforms the observed moments into treatment effect estimates. Panel (a) illustrates the relationship between the running variable and the probability of treatment, indicating 10 points with circles and labelling six of them. Panel (b) and (c) then both show the mean outcome of the treated individuals for each of these points; Panel (b) uses the running variable as the x-axis and panel (c) uses η as the x-asis. Panel (d) then shows how the Global RDD uses the information in these ten points to generate estimates of $\hat{y}_1(\eta, Z)$.

effect at a point Z by just taking the difference between the two lines at any given point; the results for this example are shown in Figure 2c.

A concern with such an observational study is that we might be concerned that individuals endogenously choose (or are chosen) whether to enroll in the treatment or not. This would cause bias in the observational study and so one could imagine trying to "debias" the observational study; of course, this begs the question of how one could do so. To see how one might do so, we can first note that in the framework presented in Section II.A, endogenous selection stems from the fact that the conditional moments potentially depend on the cost of enrollment, i.e., η. Specifically, note that we can write $\tau^*_{obs}(Z)$ as being equal to $y^*_1(\nu(Z), Z) - y^*_0(\nu(Z), Z)$, whereas the true conditional average treatment effect (CATE) is equal to $y^*_1(1, Z) - y^*_0(0, Z)$. Thus, if we can understand how $y^*_1(\eta, Z)$ and $y^*_0(\eta, Z)$ vary based on η, therefore, we could debias the observational study. Of course, this is not trivial; however, we can use the fact that we observe $y^*_1(\eta, Z^*)$ and $y^*_0(\eta, Z^*)$ at two different values of η to generate (roughly speaking) the "best guess" of how they vary based on η.

The mechanics of this can be see in Figure 2b, which shows the four observed moments (in squares). Based only these points, we will take as given for now that the "best guess" of how $y^*_1(\eta, Z^*)$ and $y^*_0(\eta, Z^*)$ vary in terms of η is a simple line through those two observed points. (We develop a Bayesian model in which this is indeed the best guess in Section III.) For notation, we will use β^*_0 to denote the slope of the implied function $\hat{y}_0(\eta, Z^*)$ and β^*_1 be the slope of the implied function $\hat{y}_1(\eta, Z^*)$, i.e.,

$$\beta^*_0 \equiv \frac{y^*_0(p_h, Z^*) - y^*_0(p_l, Z^*)}{p_h - p_l} \tag{9}$$

$$\beta^*_1 \equiv \frac{y^*_1(p_h, Z^*) - y^*_1(p_l, Z^*)}{p_h - p_l} \tag{10}$$

We could use this relationship to adjust $\hat{y}_1(\nu(Z), Z)$ and $\hat{y}_0(\nu(Z), Z)$, in hopes that it would improve the estimates that result from the observational study. The result is shown in Figure 2d. Note that the CATE jumps discontinuously at Z^* in the "Traditional Observational Study" but is smooth around the cutoff on the "Debiased Observational Study."

While this seems like a quite different approach than the Global RDD, it is in fact equivalent. We state this result formally in the remark below. Then, in Section III, we discuss the conditions under which this approach does indeed improve the resulting estimates.

Remark 1. *Let τ^*_{GRDD} be the estimate generated by the Global Regression Discontinuity Design, as defined in Section II.B, and $\tau^*_{obs}(Z)$ be the estimate generated from the traditional observational study, as defined in Equation (7). We then have:*

$$\tau^*_{GRDD}(Z) = \tau^*_{obs}(Z) - b \tag{11}$$

where b is a measure of the bias in the observational estimates. Specifically, defining β^*_0 and β^*_1 as in Equation (9) and (10), we have:

$$\begin{aligned} b &= \beta^*_0 \cdot \nu(Z) + \beta^*_1 \cdot \big(1 - \nu(Z)\big) && (12) \\ &= \xi_h \cdot \big(\tau^*_{RDD} - \tau^*_{obs}(Z^*_h)\big) + \xi_l \cdot \big(\tau^*_{RDD} - \tau^*_{obs}(Z^*_l)\big) && (13) \end{aligned}$$

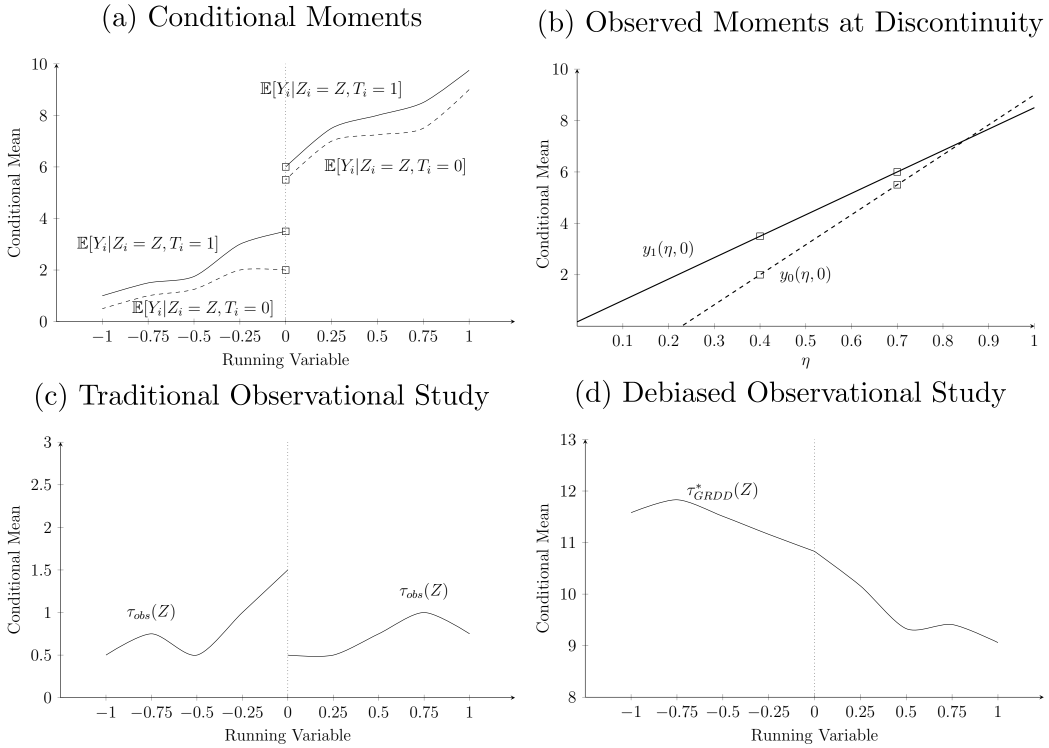
where $\xi_k \in \mathbb{R}$ is a function of p_h, p_l and $\nu(Z)$, $\tau^*_{obs}(Z^*_h) = \lim_{Z \downarrow Z^*} \tau^*_{obs}(Z)$, and $\tau^*_{obs}(Z^*_l) = \lim_{Z \uparrow Z^*} \tau^*_{obs}(Z)$.

There are two implications of the above remark. First, while we center most of the discussion around how our approach extends the traditional fuzzy RD design, you could also think of it as an approach to debias the traditional observational study design. Second, in addition to the discussion above, from which we could derive the fact that the bias estimate can be written as $\beta^*_0 \cdot \nu(Z) + \beta^*_1 \cdot \big(1 - \nu(Z)\big)$, we could instead write the bias term as a linear combination of the the traditional fuzzy RDD and observational study estimates at Z^*; see Equation 13. Thus, the Global RDD can also be thought of as a linear combination of the two traditional approaches, i.e., an observational study and a fuzzy RDD.

The intuition for this different formulation is relatively straightforward and follows from the idea that we can estimate the bias in $\tau^*_{obs}(Z)$ by comparing the fuzzy RD estimate of the LATE to the $\tau^*_{obs}(Z)$ estimates at the discontinuity. If the fuzzy RD estimates are identical to the observational estimates at the cutoff, this suggests that the observational estimates have minimal bias and so need almost no correction.[7] In contrast, if the fuzzy RD design diverges from them at the cutoff, this suggests that

[7] This is reiterates the point initially made in Battistin and Rettore (2008), which use a fuzzy RDD to validate the observational estimates; one way to view our paper is to develop an approach researches can take if their test that the observational study is unbiased, or the test proposed in Bertanha and Imbens (2020) that the local effect is generalizable to the non-compliers, is rejected.

Figure 2: Relation to Observational Studies



(a) Conditional Moments

(b) Observed Moments at Discontinuity

(c) Traditional Observational Study

(d) Debiased Observational Study

Note: This figure illustrates the relationship between the Global RDD and a traditional observational study. Panel (a) shows the values $\mathbb{E}[Y_i|Z_i = Z, T_i = 1]$ and $\mathbb{E}[Y_i|Z_i = Z, T_i = 0]$ as a function of Z. Panel (c) then transforms these moments into treatment effect estimates using the traditional observational study approach, i.e., but subtracting the two. Panel (b) shows the four observed moments at the discontinuity – shown as boxes in Panel (a) – and how these moments are used to estimate the "best guess" of how $y_1(\eta, Z^*)$ $y_1(\eta, Z^*)$ depend on η. Panel (d) then shows the results if one uses these relationship to "debias" the traditional observational study, which results in the Global RDD estimates.

the observational estimates estimates are quite biased and therefore need significant bias adjustment.

Finally, we can use the same example to highlight that it is also possible to think of the Global RDD as reflecting a natural way to extrapolate the fuzzy RDD to both the non-compiler population and away from the discontinuity. To show this, we can focus initially on extrapolating the LATE to other estimands of interest at the discontinuity. If we just use information at the discontinuity (i.e., $y_0^*(p_l, Z^*)$, $y_0^*(p_h, Z^*)$, $y_1^*(p_l, Z^*)$, and $y_1^*(p_h, Z^*)$), this is mechanically the same as extrapolating from the LATE to other estimands in an RCT context. If we use the linear approach in Brinch et al. (2017) and Kowalski (2023), therefore, we get the same result as shown in Figure 2b. After extrapolating the initial RD estimate to the non-complier population at the discontinuity, we can then use the information in the conditional moments presented in Figure 2a and a restriction that the selection does not vary away from the discontinuity to extrapolate to the population away from the discontinuity. This gives the same result shown in Figure 2d, i.e., the Global RDD estimate of the marginal treatment effects. Thus, just as the Global RDD can be thought of as a bias-adjusted observational study, it can also be thought of as an extrapolated traditional regression discontinuity design.

III Motivating the Global Regression Discontinuity Design

In the previous section, we introduced the Global RDD and showed that it can be thought of in a variety of ways: a restriction on the plausible moment functions, a linear combination of a observational study and the fuzzy RDD, or a debiased version of a traditional observational study. In this section, we start by highlighting that – at least locally – the estimator converges to the true treatment effect. Formally, we get that following theorem:

Proposition 2. *Let τ^* denote the true MTE function. Then the estimated effect on the set of compliers at the Z^* converges to the true effect on that set, i.e.:*

$$\frac{1}{p_h - p_l} \int_{p_l}^{p_h} \tau^*_{GRDD}(\eta, Z^*) d\eta = \frac{1}{p_h - p_l} \int_{p_l}^{p_h} \tau^*(\eta, Z^*) d\eta \tag{14}$$

This result is essentially the RD version of Theorem 1 of Kline and Walters (2019) and shows the Global RDD estimate of the local average treatment effect (LATE) corresponds to the true LATE even if the true $y_k^*(\eta, Z)$ functions are not in fact additively separable and linear in η. Of course, if all one was concerned about was the local average treatment effect, one could instead use a traditional fuzzy RD estimator. In contrast to a traditional fuzzy RD design, however, the global RD design also provides effect estimates away from the complier population at the cutoff. We spend the rest of this section describing different conditions under which this particular extrapolation approach can be justified.

Identified Under Weaker Assumptions: The first way we motivate the estimator is by showing that it identifies the true MTE under weaker assumptions than two of the main existing alternatives: (a) ignoring the discontinuity and relying instead on a selection-on-observables assumption and (b) the Angrist and Rokkanen method (Angrist and Rokkanen, 2015), which we discuss more below.[8]

To do so, we start by noting that if the conditional moments are indeed additively separable and linear in η then the global RDD does indeed converge to the true MTE function. This result can be stated succinctly in the following proposition:

Proposition 3. *Let y^* be the true conditional moments and suppose that $y^* \in \mathcal{Y}^{GRDD}$. Then the estimated MTE function converges to the true MTE function, i.e., $\tau^*_{GRDD} = \tau^*$.*

Absent any covariates, the assumption required for Proposition 3 is, in our opinion, relatively strong. With the addition of covariates, however, we feel like this assumption may become much more tenable. To highlight this, we show below that it relies on weaker assumptions that both a traditional observational study and the Angrist and Rokkanen method. We note explicitly, however, that the real advantage of the Angrist and Rokkanen method is that it can be used in the context of a sharp RDD, i.e., when $\nu(Z) \in \{0, 1\}$. Thus, a more precise statement would be that the Global RDD relies on weaker assumptions *in the context of a fuzzy RDD.* Finally, we acknowledge that we have not yet described how covariates can be included in the

[8] While these are not the only two methods used, the others we are aware of either focus on a marginal change in the threshold (Dong and Lewbel (2015), Cerulli et al. (2017)) or rely on additional information, such as additional covariates/measures (Mealli and Rampichini (2012), Wing and Cook (2013), Rokkanen (2015)) or multiple discontinuities (Cattaneo et al. (2021), Bertanha (2020)). We discuss our method in a context with multiple discontinuities in Section B.

Global RDD; we leave the formalization of this to Appendix A, but note here that in the result below we are implicitly including the same set of (exogenous) covariates in all three estimators.

Proposition 4. *The assumptions under which the Global RDD converges to the true treatment effect function are weaker than those under which the observational study does. Similarly, the conditions under which the Global RDD converges to the true average treatment effect on a complier population are weaker than the conditions under which the Angrist and Rokkanen (2015) estimator does.*

While we leave the formal proof to Appendix A, we highlight here some nice similarity in the intuition behind why the global RDD relaxes the assumptions required in the traditional observational study and those required in the Angrist and Rokkanen method. In particular, the traditional observational study assumes that there is no endgoenous selection into the treatment which, in the MTE formulation, amounts to the assumption that $y_k(\eta, Z)$ does not depend on η. The Angrist and Rokkanen method, in contrast, allows for endogenous selection into the treatment but assumes that the moments do not depend on the running variable; in the MTE formulation, this amounts to the assumption that $y_k(\eta, Z)$ does not depend on Z. The Global RDD therefore relies on weaker assumptions since it allows for the conditional moments to depend on both η and Z. As described in Section II.C, however, the Global RDD does put restrictions on how the conditional moments can vary based on η and Z and so we next turn our attention to ways in which the estimator can be motivated under weaker assumptions than what is required in Proposition 3.

Optimal Estimator in a Bayesian Model: Even though the Global RDD converges to the true MTE under weaker assumptions than the main alternative approaches, we still may not be comfortable with the assumptions required to identify the true MTE functions. Motivated by the researchers' own experience, we therefore consider in the section what happens when the researcher is not willing to assume that the functions are necessarily additively separable and linear in η, but also does not have a strong intuition on how they deviate from it. We then show that under certain conditions, the Global RDD may still correspond to the researchers' "best guess" at the MTE function given the observed data, even if it's not necessarily the true MTE function.

To do so, we introduce a model in which the conditional moment functions themselves are generated randomly according to some probability measure. A natural way to think of this model – and one that is consistent with the motivation above – is as a Bayesian hierarchical model, in which case the probability measure over the conditional moment functions serves as a Bayesian prior. An alternative is to view the results as measuring the expected performance of the estimator over a range of empirical scenarios; in this interpretation, the probability measure is defined by the range of empirical scenarios that is considered.

To more formally introduce the model, we will let $\mathcal{Y} = \mathcal{Y}_0 \times \mathcal{Y}_1$ be the set of potential conditional moment functions. We will then assume that the functions in $y_k \in \mathcal{Y}_k$ for $k \in \{0, 1\}$ are distributed according to a Gaussian process (GP) as follows:

$$y_k(\eta, Z) = \alpha_k + \beta_k \eta + \gamma_k Z + \tilde{y}_k(\eta, Z) \tag{15}$$

$$\tilde{y}_k(\eta, Z) \sim \mathcal{GP}(0, C_k) \tag{16}$$

$$[\alpha_k, \beta_k, \gamma_k]' \sim N(0, \sigma^2 I) \text{ with } \sigma^2 \to \infty \tag{17}$$

where $\mathcal{GP}(0, C_k)$ corresponds to a mean-zero Gaussian process with covariance function C_k and I corresponds to the identity matrix.[9]

For the interested reader, Rasmussen and Williams (2006) provides an excellent introduction to Gaussian processes. Here, we will simply note two important points about Gaussian processes in general and the one specified in Equations (15) - (17) in particular. First, it is precisely the covariance function C_k that determines the implied prior distribution over functions $y_k \in \mathcal{Y}_k$. We will not focus on the particulars of how it does so here, but a wide range of priors can be achieved by varying C_k. Second, by considering the case in which $\sigma^2 \to \infty$ we use an uninformative prior limit for the linear terms, in which all linear functions of η and Z are considered equally likely.

Finally, we note that the model does not require fundamental changes to the one presented in Section II.A. Instead of referring to y_1^* and y_0^* as the true conditional moment functions, we simply view them as a single realization of the Gaussian process.[10] Furthermore, note that the Gaussian process defined in Equations (15) - (17)

[9] There have been papers, e.g., Branson et al. (2019), that propose using a Gaussian process in the context of a RDD. Others, e.g., Chib and Jacobi (2016) and Chib et al. (2023), develop Bayesian approaches to estimate RDDs. They do so in the context of estimation rather than, in our case, as a way to interpret and motivate a particular extrapolation away from the discontinuity.

[10] We will ensure that every y_1^* and y_0^* generated by the model specified in Equations (15) - (17) satisfies the assumptions specified in Section II.A by assuming that the covariance functions C_k are such that the Gaussian process is sample continuous, i.e., that every y_1^* and y_0^* generated by the model is a continuous function. This can be done by assuming that the covariance functions

along with the linear transformation defined in Equation (3) together imply a prior distribution over the MTE functions $\tau(\eta, Z)$ and we can similarly use τ^* to be a particular realization of the GP.

As an aside, one can combine a choice of prior (i.e., a choice of C_k) with the observed data to generate Bayesian posteriors of the y_k functions, and hence Bayesian posteriors of the MTE function. That is the approach taken in the context of an RCT in Opper (2023), which discusses how a similar model presented above can be used to generate posterior distributions of the average treatment effect (ATE), the average treatment effect on the treated (ATT) and other potentially non-identifed parameters of interest. However, even if we restrict ourselves to (for example) the commonly used squared-exponential covariance functions, the identification of the hyperparameters is even more challenging in the RD setting than in the RCT setting studied in Opper (2023).[11] We therefore opt instead to study in this paper how the Global RDD compares to other approaches under all possible GPs, within the flexible framework presented in Equations (15) - (17). That is, are there ways to motivate the Global RDD that do not depend on the specification of C_k?

Much of the intuition for our analysis here stems from the following proposition, which states that – *regardless of the choice of* C_k – the Global RDD corresponds to our best guess of the MTE function based on the moments we observe at the discontinuity and one additional point. Formally, letting $\mathcal{D}(A)$ denote the observed conditional moments at points $Z \in A$, the proposition is as follows:

Proposition 5. *For any choice of C_0 and C_1 and any point $\tilde{Z} \neq Z^*$, we get that:*

$$\tau^*_{GRDD}(\eta, Z) = \mathbb{E}\big[\tau^*(\eta, Z)|\mathcal{D}(\{\tilde{Z}, Z^*\})\big] \tag{18}$$

for every η and $\in \{\tilde{Z}, Z^*\}$.

This also implies that – again regardless of the choice of C_k – the bias correction used in the Global RDD (as described in Section II.C) is the best guess of the bias in the observational study based on the four moments observed at this discontinuity, or that:

Lipschitz continuous, although that is not a necessary condition. Finally, we could allow, as in Opper (2023), the realizations of y_0^* and y_1^* to be correlated, but do not to do here for simplicity.

[11] This additional challenge stems from the fact that we need to specify hyperparameters that both govern both the direct effects of η and Z on $y_k(\eta, Z)$ as well as the interaction between the two.

Proposition 6. *Define to $b^*_{obs}(\eta, Z)$ be the bias in the observational study, i.e.,*

$$b^*_{obs}(\eta, Z) = \tau^*_{obs}(Z) - \tau^*(\eta, Z) \tag{19}$$

where $\tau^*_{obs}(Z)$ is defined in Equation (7). Then:

$$\tau^*_{GRDD}(\eta, Z) = \tau^*_{obs}(Z) - \mathbb{E}\big[b^*_{obs}(\eta, Z)\big|\mathcal{D}(\{Z, Z^*\})\big] \tag{20}$$

for any (η, Z) and any choice of C_0 and C_1.

These result seems quite promising; however, we have assumed (via Assumption 5) that we observe data at all points $Z \in \mathbf{Z}$ and not just at two points and (similarly) want to adjust the observational study using the best guess of the bias using all the data. Unfortunately, Proposition 5 does not extend to show that $\tau^*_{GRDD}(\eta, Z) = \mathbb{E}\big[\tau^*(\eta, Z)|\mathcal{D}(\mathbf{Z})\big]$ or that $\mathbb{E}[b^*_{obs}(\eta, Z)|\mathcal{D}(\mathbf{Z})]$ regardless of the covariance functions. The reason is that it is the covariance functions that determine whether we ascribe deviations from linearity in the observed $y_k(\nu(Z), Z)$ to: (a) non-linearities in the relationship between Z and y_k; (b) non-linearities in the relationship between η and y_k; or (c) interactions between η and Z. The Global RDD method instead ascribes all such deviations to non-linearities in the relationship between Z and y_k.

There are, however, some conditions under which we can indeed interpret the Global RDD as the mean posterior conditional on all the observed data. In particular, the Global RDD can be thought of as the best guess of the MTE function if we are willing to assume that selection into treatment does not vary based on Z (but that it is not necessarily linear in η) and the treatment thresholds do not vary away from the discontinuity. Formally, this is stated in the following proposition:

Proposition 7. *Suppose that $C_k((\eta, Z), (\eta', Z')) = C_{k,\eta}(\eta, \eta') + C_{k,Z}(Z, Z')$ for both $k \in \{0, 1\}$ and that:*

$$\nu(Z) = \begin{cases} p_l \text{ if } Z < Z^* \\ p_h \text{ if } Z > Z^* \end{cases} \tag{21}$$

Then for any choice of $C_{k,\eta}$ and $C_{k,Z}$, we get that:

$$\tau^*_{GRDD}(\eta, Z) = \mathbb{E}\big[\tau^*(\eta, Z)|\mathcal{D}(\mathbf{Z})\big] \tag{22}$$

for every (η, Z).

The two additional conditions specified in Proposition 7 – i.e., that $\nu(Z)$ is a step-function and that the functions $y_k(\eta, Z)$ are separable – are particularly interesting because in a small neighborhood around the discontinuity they are guaranteed to be (nearly) satisfied. This implies that the Global RDD is (nearly) the mean posterior as long as we restrict ourselves to a small enough neighborhood around the discontinuity.[12]

The fact that the Global RDD is, under some conditions, the mean posterior of the MTE function is also important because it implies that the Global RDD is the optimal estimator in a Bayesian decision theory model. To state this formally, given a realization of the modified GP and its implied MTE function τ^* as well as an estimate of the MTE function $\hat{\tau}$, we then define the loss as the mean-squared error, i.e.,:[13]

$$l(\hat{\tau}, \tau^*) = \Big(\tau^*(\eta, Z) - \hat{\tau}(\eta, Z)\Big)^2 \tag{23}$$

This loss clearly depends on the observed data (which determines $\hat{\tau}$) and the realization of the modified GP (which determines τ^*). It is natural, therefore, to evaluate the performance of an estimator using the expected loss, where the expectation is taken over realization of the modified GP and the observed data. Specifically, we can evaluate any specific estimator by the value of its expected loss, i.e.,:

$$\mathscr{L} = \mathbb{E}[l(\hat{\tau}^*, \tau^*)] \tag{24}$$

with the value of $\mathscr{L}$ clearly depending on both the formulation of the modified GP (i.e., the specification of the covariance functions) and the way the observed data is transformed into estimates of the MTEs (i.e., the estimator used).

From the fact that mean posterior under the conditions in Proposition 7, we also get that it corresponds to the estimator that minimizes the expected loss, i.e., that minimizes $\mathscr{L}$. We can also use this framework to compare the expected loss of the Global RDD to alternative approaches. In particular, we can show that regardless

[12] The fact that we can conclude from the fact that assumptions are "nearly" satisfied that the Global RDD is "nearly" the mean posterior stems from the the fact that the mean posterior can be thought of as the maximizer of the posterior probability distribution, which depends continuously on the functions ν and C.

[13] While we focus on the mean-squared error, the results generally apply to any symmetric loss function. Furthermore, while we focus on the loss at a specific value of (η, Z), we could also follow Mogstad et al. (2018) and instead specify that the researcher is interested in $\Gamma(\tau) = \int_{\mathbf{Z}} \int_0^1 \tau(\eta, Z)\omega(\eta, Z)$ for some weighting scheme $\omega(\eta, Z)$.

of the choice of C_0 and C_1, the Global RDD results in lower expected loss than a traditional observational study. Furthermore, as long as the modified GP allows for sufficient possibility that the treatment effect varies with the running variable, the Global RDD also results is lower expected loss than a traditional fuzzy RD design. See Propositions 9 and 10 in Appendix A for the formalization of these statements and their proofs.

IV Estimation Approach

So far, we have assumed that the researchers observe the true conditional moment functions $\mathbb{E}[Y_i|T_i = k, Z_i = Z]$ and $\mathbb{E}[T_i|Z_i = Z]$ for all $Z \neq Z^*$ and $k \in \{0,1\}$. In practice, of course, these moments need to be estimated. In this section, we first discuss what assumptions regarding the data generating process can replace the assumption that the true conditional moment functions are observed. We then outline our estimation approach and show that, under the new assumptions, the resulting estimate converges to τ^*_{GRDD}.

The four additional assumptions that collectively replace Assumption 5 are listed below:

Assumption 6. *$\mathbb{E}[\mu_i^2|\eta_i = \eta, Z_i = Z] < \infty$ and $\mathbb{E}[\tau_i^2|\eta_i = \eta, Z_i = Z] < \infty$ for all $\eta \in (0,1)$ and $Z \in \mathbf{Z} \equiv (\underline{Z}, \overline{Z})$.*

Assumption 7. *Z_i is continuously distributed over $\mathbf{Z}$ with a strictly positive distribution function.*

Assumption 8. *The parameter space for $\nu(Z)$ is $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$ where:*

$$\nu(Z_i) = \begin{cases} \nu_1(Z) & \text{if } Z < Z^* \\ \nu_2(Z) & \text{if } Z > Z^* \end{cases}$$

and where $\nu_k \in \mathcal{V}_k$. Let $\mathcal{W}_{2,2}$ be the Sobolov space of functions $f : \mathbf{Z} \to \mathbb{R}$ and $||f||_{\mathcal{W}_{22}}$ be its norm, as defined in Freyberger and Masten (2019). Then $\mathcal{V}_k = \{\nu_k \in \mathcal{W}_{22} : ||\nu_k||_{\mathcal{W}_{22}} < V_k\}$ for some constant V_k. Furthermore, we will assume that the true function $\nu^* \in \mathcal{V}$.

Assumption 9. *Let $\mathcal{W}_{2,2}$ be the Sobolov space of functions $f : \mathbf{Z} \to \mathbb{R}$ and $||f||_{\mathcal{W}_{22}}$ be its norm, as defined in Freyberger and Masten (2019). Then $\mathcal{Y}_k^{GRDD} = \{y_k =$*

$\beta_k\eta + \gamma_k(Z) : \beta_k \in [\underline{K}, \overline{K}]$ and $\gamma_k \in \mathcal{W}_{22} : ||\gamma_k||_{\mathcal{W}_{22}} < Y_k\}$ for some constants $\underline{K}$, $\overline{K}$, and Y_k. Furthermore, we will assume that there exists $y_k^* \in \mathcal{Y}_k^{GRDD}$ such that $y_k^*(\nu(Z), Z) = \mathbb{E}[Y_i|T_i = k, Z_i = Z]$ for all $Z \in \mathbf{Z}$.

All four assumptions are, in our opinion, relatively benign assumptions and reflect common assumptions made in econometrics literature. Assumption 6 simply states that the individuals' outcomes have finite variance at every point, which allow us to use the standard asymptotic methods. Assumption 7 ensures that asymptotically there will be a large number of observations arbitrarily close to each point $Z \in \mathbf{Z}$.

Assumptions 8 and 9 state that the parameter space is compact for both ν and y, which helps ensure that our non-parametric estimation approaches converge. As written, this assumptions permits us to consider uniform convergence and both could be relaxed if one was only interested in pointwise convergence. Note also that by assuming that the true functions fall within the parameter space, these also capture the assumption that the true functions are smooth.

We next formally define our estimator as follows:

Global Regression Discontinuity Design. *Our proposed estimator consists of three steps:*

1. *Estimate $\nu(Z)$ as follows:*

$$\hat{\nu} = \arg\min_{\nu \in \mathcal{V}} \left\{ \sum_{\forall i} \Big(T_i - \mathbf{1}(Z_i > Z^*)\cdot\nu(Z_i) - \mathbf{1}(Z_i < Z^*)\cdot\nu(Z_i)\Big)^2 + \lambda_\nu \int \Big(\nu''(Z)\Big)^2 dZ \right\}$$

2. *Estimate y as follows:*

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}, \hat{\delta} = \arg\min_{\beta_0, \beta_1, \gamma, \delta} \left\{ \sum_{\forall i} \left(Y_i - \Big(\beta_0 \cdot \hat{\nu}(Z_i) + \gamma(Z_i) + T_i \cdot \big((\beta_1 - \beta_0)\cdot\hat{\nu}(Z_i) + \delta(Z_i)\big)\Big)\right)^2 \right.$$
$$\left. + \lambda_\gamma \int \Big(\gamma''(Z)\Big)^2 dz + \lambda_\delta \int \Big(\delta''(Z)\Big)^2 dZ \right\}$$

and then:

$$\hat{y_0}(\eta, Z) = \hat{\beta}_0\eta + \hat{\gamma}(Z_i)$$
$$\hat{y_1}(\eta, Z) = \hat{\beta}_1\eta + \hat{\gamma}(Z_i) + \hat{\delta}(Z_i)$$

3. *Estimate τ using $\hat{y}$ as follows:*

$$\hat{\tau}(\eta, Z) = T\big(\hat{y}_0, \hat{y}_1\big)$$

where:

$$\begin{aligned} T(y_0, y_1) &= y_1 - y_0 + \eta\frac{\partial y_1}{\partial \eta} + (1-\eta)\frac{\partial y_0}{\partial \eta} \\ &= \hat{\delta}(Z) + (\hat{\beta}_1 - \hat{\beta}_0)\eta + \hat{\beta}_1\eta + \hat{\beta}_0(1-\eta) \\ &= \hat{\delta}(Z) + \hat{\beta}_0 + 2(\hat{\beta}_1 - \hat{\beta}_0)\cdot\eta \end{aligned}$$

We next show that with enough data the estimator defined above will result in an estimate of the marginal treatment effect (MTE) function that becomes arbitrarily close to the Global RDD estimator defined and analyzed in the previous sections, i.e., to τ^*_{GRDD}. The formal theorem is below, with the proof in Appendix A:

Proposition 8. *Let $\hat{\tau}_{GRDD}$ be the estimate generated by the Global Regression Discontinuity Design, as defined above, and $\hat{\tau}^*_{GRDD}$ be the MTE function defined in Definition 1. Then given Assumptions 1-4 and 6-9 we get that:*

$$\hat{\tau}_{GRDD} \xrightarrow{p} \tau^*_{GRDD} \tag{25}$$

Note that choice of norms in Assumption 8 and 9 implies that the covergence occurs in a uniform sense, rather than a pointwise sense. Specifically, $\hat{\tau} \xrightarrow{p} \hat{\tau}^*$ means that for all $\epsilon, \delta > 0$ there exists an $\overline{N}$ such that $\forall n \geq \overline{N}$ we have that $\mathbb{P}\big(sup\{|\hat{\tau}_n(\eta, Z) - \hat{\tau}^*(\eta, Z)| : (\eta, Z) \in (0,1) \times \mathbf{Z}\} > \epsilon\big) < \delta$.

Finally, as mentioned in Section III, the assumption required for τ^*_{GRDD} to be equal to the true MTE function will, depending on the context, potentially be more believable when one conditions on a set of exogenous covariates. In one sense, extending the model to condition on a set of covariates is straightforward. If one extends Assumption 5 to be that we observe $\mathbb{E}[T_i|Z_i = Z, X_i = X]$, $\mathbb{E}[Y_i|T_i = 0, Z_i = Z, X_i = X]$, and $\mathbb{E}[Y_i|T_i = 1, Z_i = Z, X_i = X]$ at every point $Z \neq Z^*$ and X and similarly extends the other assumptions to also be conditional on $X_i = X$, then nothing about the method or results would need to change. In practice, however, estimating the conditional moments non-parametrically, as we do above, quickly gets challenging as the number of covariates increases. Implementing the method therefore will likely require some additional restrictions.

This raises the intriguing possibility that including covariates could lead to better identification of the MTE functions. For example, if the size of the discontinuity in $\nu(Z, X)$ at Z^* differed depending on X and we keep the restriction that $y_k(\eta, Z, X)$ is additively separable, we could relax the restriction that it is linear in η. This is because we now see multiple conditional moments which we can use to identify the dependence of $y_k(\eta, Z, X)$ on η.[14]

We view this possibility, and in particular the best way to extend the Bayesian model outlined in Section III to account for additional covariates, to be an area ripe for further exploration. At this point, however, we leave it for further exploration and instead have implemented an approach in the accompanying R package that accounts for additional covariates in similar way as other approaches that include covariates into a regression discontinuity design. Specifically, we account for a vector of additional exogenous covariates X_i by adjusting Step 2 of the Global Regression Discontinuity Design to be that we estimate y as follows:

$$\begin{aligned}\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}, \hat{\delta}, \hat{\xi} = \arg\min_{\beta_0,\beta_1,\gamma,\delta} \Bigg\{ \sum_{\forall i} \Bigg(Y_i - \Big(\beta_0 \cdot \hat{\nu}(Z_i) + \gamma(Z_i) + T_i \cdot \big((\beta_1 - \beta_0) \cdot \hat{\nu}(Z_i) + \delta(Z_i)\big) + \xi X_i \Big)\Bigg)^2 \\ + \lambda_\gamma \int \Big(\gamma''(Z)\Big)^2 dz + \lambda_\delta \int \Big(\delta''(Z)\Big)^2 dZ \Bigg\}\end{aligned}$$

and where then:

$$\begin{aligned}\hat{y_0}(\eta, Z) &= \hat{\beta}_0 \eta + \hat{\gamma}(Z_i) + \hat{\xi} X_i \\ \hat{y_1}(\eta, Z) &= \hat{\beta}_1 \eta + \hat{\gamma}(Z_i) + \hat{\delta}(Z_i) + \hat{\xi} X_i\end{aligned}$$

To implement the Global RDD, we use the MGCV package (Wood (2017)) in R which automatically chooses the smoothing parameters λ_ν, λ_γ, and λ_δ. We use a Bayesian bootstrap procedure in which we randomly specify weights, drawn from a Dirilecht distribution, and then repeat the Global Regression Discontinuity Design procedure specified above to account for the estimation of ν when computing standard errors. See https://github.com/isaacopper/GlobalRDD for the R package that

[14] The resulting analysis is similar to the previously cited Brinch et al. (2017), which discusses identification in a non-RD context.

implements these packages.

V Empirical Application: Grade Retention Policies

In this section, we present an application of our estimator in education policy: a field where fuzzy RD design has become more popular with the increasing use of student test score cutoffs (or performance index cutoffs based on student test scores) to identify eligibility for educational interventions. In particular, we explore the broader effects of test-based retention policies. As we detail below, there is extensive literature examining the effects of grade retention on student outcomes using fuzzy RD designs; however, these estimated effects often apply only to compliers (i.e., students not exempt from retention) right below retention cutoffs. In this exercise, we ask whether these effects differ for exempt students and for lower-performing students identified for retention.

V.A Policy Background and Data

Calls to end social promotion in schools in the 1990s and an increased popularity of educational accountability and standardized testing led to test-based retention policies in many states and school districts in the United States over the past three decades. Perhaps the most influential of these policies has been Florida's third grade retention policy, which was enacted in 2002 and provided the blueprint for others nationwide. This policy requires students who score in the lowest achievement level on statewide reading test to repeat third grade and receive instructional support (e.g., additional instruction time in reading, being assigned to highly effective teachers).

There are several "good cause exemptions" that allow students to be promoted to the fourth grade despite failing to score at the Level 2 benchmark or above. In particular, students in the lowest achievement level in reading can be promoted to fourth grade (1) if they have been in the English learner program for less than two years; (2) if they have certain disabilities and have been already retained once until third grade; (3) if they have received intensive reading remediation for two years and have already been retained twice between kindergarten and third grade; (4) if they demonstrate that they are reading at a level equal to or above a Level 2 on the statewide reading test by performing at an acceptable level on an alternative standardized reading assessment approved by the State Board of Education; or (5) if they demonstrate proficiency through a teacher-developed portfolio. Despite these exemptions, the policy has affected a significant share of third graders in the state: in the first year of the policy, 21 percent of third graders were flagged for retention (i.e., scored below the retention cutoff) and 15 percent had to repeat third grade (Licalsi et al. (2019)). Among those flagged for retention, one-third received an exemption and were promoted to fourth grade. While retention rates gradually declined partly due to improvements in reading achievement and the increase in exemption rates, they remained sizable with roughly 10 percent of the third graders being retained in 2021-22 school year.

Several studies have examined the effects of being retained (and receiving instructional support) under Florida's retention policy on student outcomes using the discontinuity in retention likelihood and RD designs (Greene and Winters (2007), Winters and Greene (2012), Özek (2015), Schwerdt et al. (2017), Figlio and Özek (2020)). The overarching conclusion is that retained students outperform their same-age peers in the short term (one to three years), these achievement gains fade out over time. That said, retained students under Florida's retention policy significantly outperform their promoted peers when they reach the same grade level, and are also less likely to be retained in a later grade. While providing compelling evidence, by using traditional RD designs these papers all focus on the complier population at the discontinuity. In this paper, we use the proposed estimator to determine how these benefits differ for students away from the cutoff and for students who were promoted to fourth grade using exemptions.

To address these questions, we use student-level administrative data from a large urban school district (LUSD) in Florida. In our analysis, we use students who entered third grade for the first time between 2005-06 and 2010-11 school years and follow them until 8th grade. Roughly 17 percent of these students were flagged for retention and of those identified for retention, 38 percent were retained, corresponding to 7 percent of the third graders in these cohorts. Of those who were not flagged for retention, a small number of students ($\tilde{1}$ percent) were retained regardless and so there is two-sided non-compliance in this setting. Our main outcomes of interest are standardized reading scores in grades 4 through 8.[15]

[15] In the analysis that follows, we use a same-grade comparison: That is, we compare the test

V.B Results

Figures 3 and 4 (along with Table 1) present the estimated effects for exempt and non-exempt students around and away from the retention cutoff in different ways. The overarching conclusion from this analysis is that the impact of retaining students is largest for those with the lower third grade reading scores and for those who – conditional on their third-grade reading score – are most likely to be retained.

For example, Figure 3 shows how the conditional average treatment effect on the treated individuals depends on third grade reading scores. Specifically, the solid line shows how the estimated effect – formally $\mathbb{E}[\tau_i|Z_i = Z, T_i = 1]$ – depends on the value of the running variable (third grade reading scale scores centered at the retention cutoff). The dashed lines indicate the 95 percent point-wise confidence interval and were using a Bayesian bootstrap procedure, in which we repeatedly (n = 100) drew weights for each student from a Dirichlet distribution and estimate $\hat{\tau}$ using the procedure defined in Section II.B. The dashed lines then illustrate the range of these estimates.[16]

The results in Figure 3 suggest that the positive effects of retention on fourth grade reading scores monotonically decline with students' baseline reading achievement. At the cutoff, we find that retention increases fourth grade reading scores by roughly 0.9σ, which is consistent with the effect sizes found in the previous literature (Schwerdt et al. (2017), Figlio and Özek (2020)). This benefit grows to 1.2σ for students whose third grade reading scores fell 25 points below the cutoff, and to 1.4σ for students 50 points below the cutoff. In contrast, the positive effects decline to 0.8σ for students 25 points above the cutoff and to 0.6σ for those 50 points above. Since most students who are retained are below the cutoff, these findings suggest that the LATE estimates presented in prior RD studies in this context significantly underestimate the overall benefits of retention in the short term.

It is also clear from Figure 3 that at the discontinuity, the effect on the treated individuals jumps. This stems from the fact that, by construction, the characteristics

scores of retained and promoted students when they reach the same grade level. Another approach commonly used in the grade retention literature is to compare the test scores of treated and comparison students in years following the treatment (i.e., same-age comparison). We prefer the former approach as we see additional time provided to retained students as part of the treatment. That said, we also conducted a same-age comparison (results available upon request) and the main conclusions remain unchanged.

[16] We used a Bayesian bootstrap instead of a traditional bootstrap to ensure that in every iteration there was two-sided imperfect compliance at the discontinuity.

of the treated population discontinuously change at the threshold. We illustrate the effect of this more directly in Figure 4, which illustrates how $\hat{\tau}(\eta, Z)$ varies by both Z and η. In this exercise, η_i can be interpreted as "promotion likelihood": a student is retained if and only if their η_i falls below a given cutoff. In other words, effect estimates for higher values of η_i indicate the retention effect for students who are least likely to be retained and vice versa. In this graph, each line corresponds to a set of (η_i, Z_i) values with the same estimated effect. There are two important takeaways from this figure.

First, consistent with Figure 3, the estimated effect declines as students' baseline reading achievement increases (moving from left to right). Second, we also observe that students who are less likely to receive an exemption and be promoted to fourth grade benefit significantly more from retention. For example, at the retention cutoff, the average effect for students who are least likely to be retained (η_i=1) is roughly 0.3σ where the average effect for those most likely to be retained (η_i=0) is 1σ. This finding suggest that the exemptions to the retention rule incorporated into Florida's policy indeed identify students who are least likely to benefit from retention. That said, exempt students with lower baseline achievement would still benefit from retention: the effect of retention on students 25 points below the cutoff who are most likely to receive an exemption is nearly 0.7σ while the effect for exempt students 50 points below the cutoff is roughly equivalent to the effect of retention for non-exempt students at the cutoff.

Table 1 extends this analysis to reading scores in grades 4 through 8 under different scenarios for treatment assignment. In the first column, we present the treatment effects under optimal treatment assignment (i.e., keeping the retention rate constant, yet assigning the individuals who would benefit most from retention). The second column presents the average treatment effect under the realized treatment assignment (average treatment effect on the treated or ATT); the third column gives the estimated effect on the complier population at the threshold (or LATE); the fourth gives the average treatment effect if students were randomly retained (overall average treatment effect on the treated or ATE); and the last column provides the average treatment effect if those who are not retained under the realized treatment assignment were retained (average treatment effect on the controls or ATC).

The results suggest that the realized assignment is nearly equivalent to the optimal assignment. In particular, average treatment effects under realized assignment are larger than 77 percent of the average treatment effects under optimal assignment in all cases. It also suggests that the ATT is larger than the LATE, implying that the policy increases test scores of those retained by more than has been shown in previous studies which use an RD approach to identify the LATE. However, the results in the last column shows that expanding the program would have minimal effects in the years after the student was retained and that these effects fade-out completely by sixth-grade. This suggests that Florida's policy is quite successful in identifying students most likely to benefit from retention.

VI Conclusion

The trade-off between internal and external validity is a common issue in causal inference. In the context of RD design, this trade-off manifests itself in two ways. First, the RD estimates obtained using traditional methods only apply to individuals identified for treatment within a small bandwidth around the treatment cutoff. Second, in many RD applications, treatment assignment is fuzzy: that is, being on the treatment side of the cutoff does not fully determine treatment status due to non-compliance or policy-dictated exemptions. In those settings, it is hard to generalize traditional RD estimates to non-compliers. That said, understanding treatment effects beyond compliers around the treatment cutoff is critical from a public policy perspective in many settings for several reasons.

In this study, we propose a new method for use in fuzzy RD settings, which we call the Global Regression Discontinuity Design, to address this issue. The estimator can be thought of either as a bias-adjusted observational study or an extrapolation of the traditional fuzzy regression discontinuity estimate (first to non-compliers at the cutoff and then to individuals away from the cutoff). We then show that it can be motivated in both a frequentest framework (in that it is consistent under weaker assumptions than existing approaches) or in a Bayesian framework (in that can be considered the posterior mean given the observed conditional moments under more flexible conditions).

We then present an application of this method in education policy. In particular, we examine the broader effects of early grade retention policies, which often require students to score above a predetermined threshold on third-grade reading tests to be promoted to fourth grade, on student outcomes using student-level data from

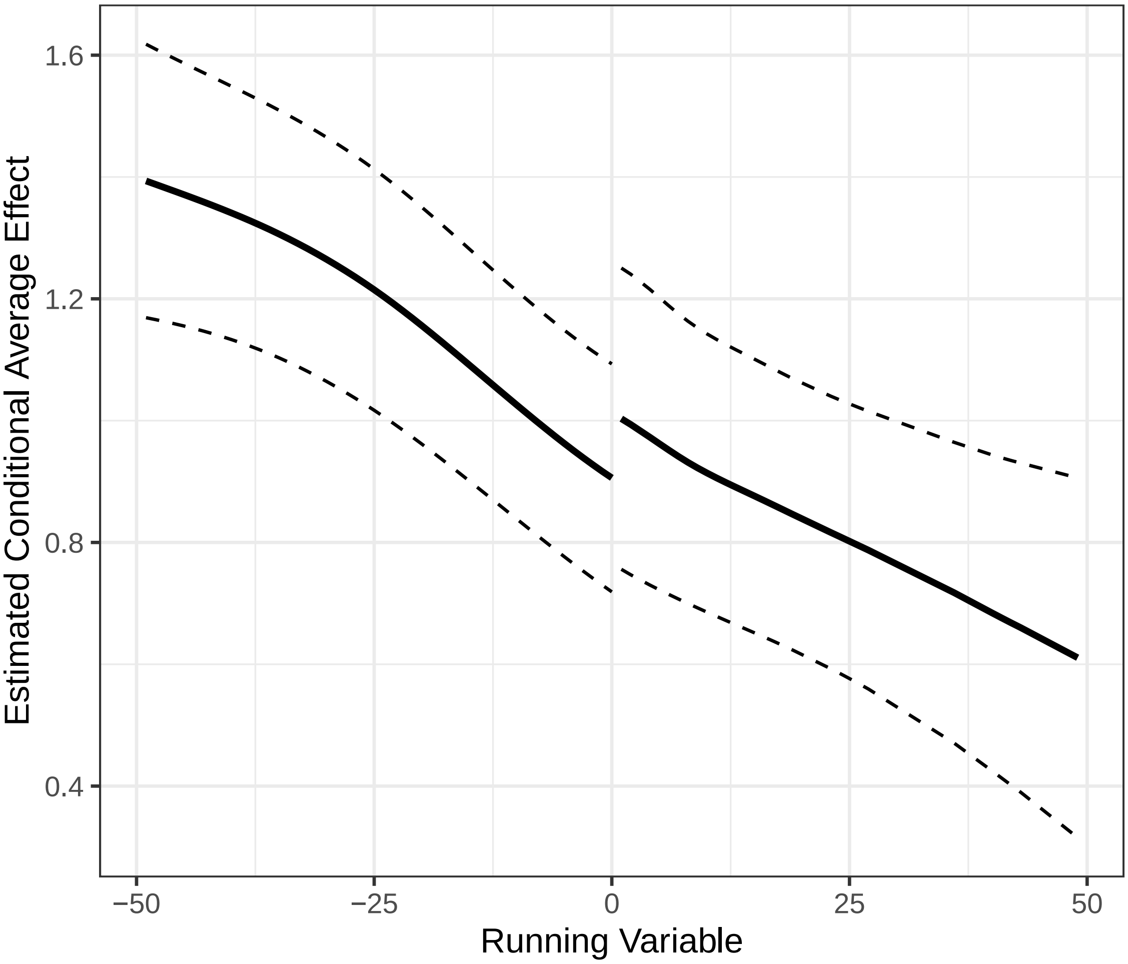
Florida. Several prior students have addressed this question using traditional RD designs and found significant benefits. Here, we ask how these benefits differ for lower-performing students away from the cutoff and for low-performing students who were promoted using exemptions. We find that the positive effects of retention are larger for students with lower baseline reading achievement and smaller for student exempt from retention. Our findings also suggest that retaining more students, by either increasing the threshold or removing exemptions, would have limited effect on the newly retained students.

Finally, we conclude by highlighting that the marginal treatment effect representation of the fuzzy RDD settings provides a natural framework for researchers to consider ways of extending the method presented above to slightly different contexts. While we focus on the most simple design here, we sketch in Appendix B how the model can be extended to handle multiple discontinuities, provide an alternative tests for the external validity of the traditional RDD estimates, handle cases of one-sided non-compliance, and be used to improve the precision of the fuzzy RDD estimates.

VII Graphs and Tables

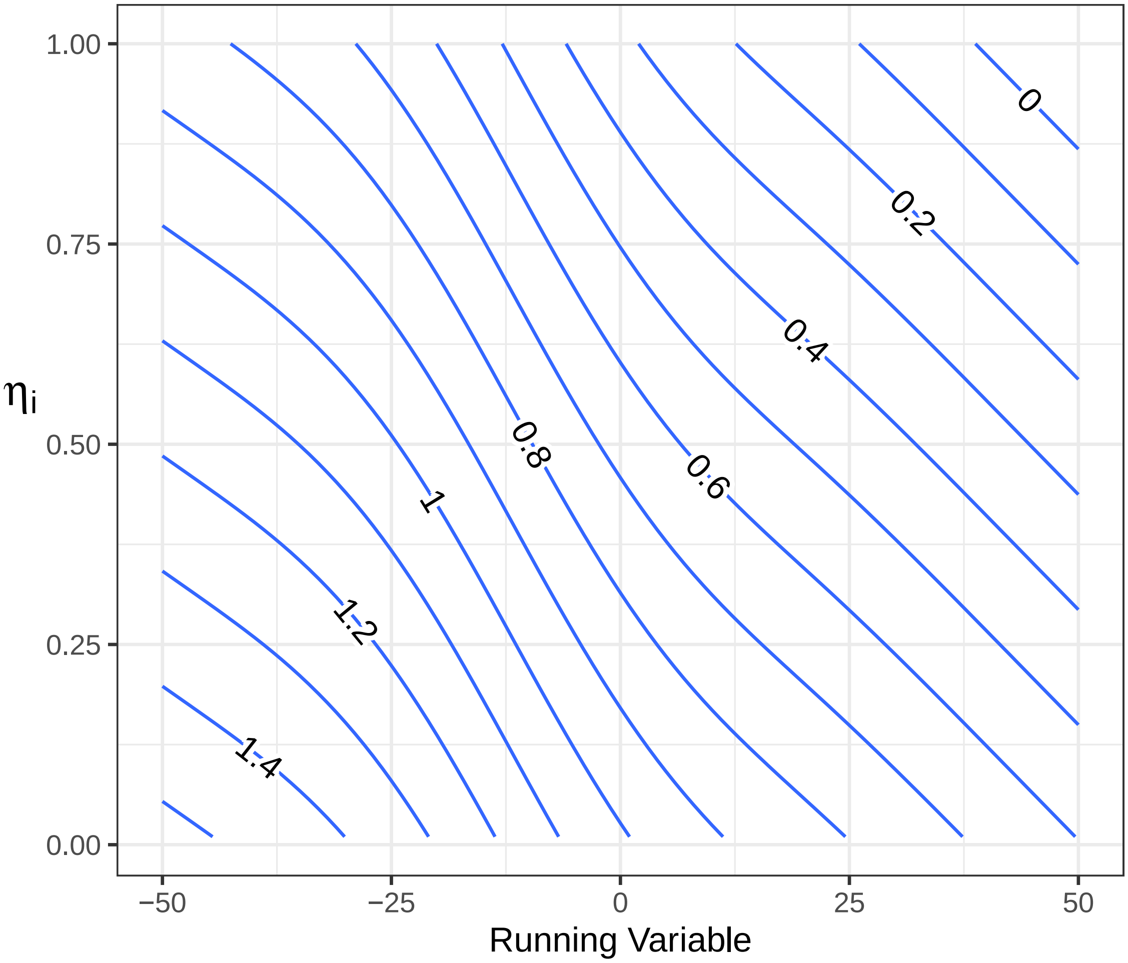
VII.A Graphs

Figure 3: Average Treatment Effect on the Treated



Note: The figure plots how the estimated the conditional average treatment effect on the treated varies with the running variable. Specifically, the solid line shows the estimated $\hat{\mathbb{E}}[\tau_i | Z_i = Z, T_i = 1]$ and the dashed lines indicated the 95% confidence interval, estimated via a Bayesian bootstrap with school-level clustering.

Figure 4: Estimates of $\tau(\eta, Z)$



Note: The figure illustrates how $\hat{\tau}(\eta, Z) = \mathbb{E}[\tau_i | \eta_i = \eta, Z_i = Z]$ varies with both Z and η. Each line corresponds to a set of (η, Z) values with the same value of $\hat{\tau}(\eta, Z)$. Roughly speaking, η_i is a latent variable that serves as a measure of how likely an individual is to enroll in the treatment; individuals' with low values of η_i are more likely to enroll than individuals with high values and so it is sometimes referred to as the "latent cost" of enrolling. See Section II.A for the formal definition.

VII.B Tables

Table 1: Average Effect with Different Treatment Assignments

	Optimal Assignment	Realized Assignment (ATT)	Local Effect (LATE)	Random Assignment (ATE)	Program Expansion (ATC)
Grade 4	1.36	1.11	0.90	0.50	0.42
	(0.42)	(0.11)	(0.12)	(0.47)	(0.53)
Grade 5	0.84	0.74	0.62	0.37	0.31
	(0.15)	(0.11)	(0.11)	(0.27)	(0.33)
Grade 6	0.69	0.61	0.53	0.07	−0.00
	(0.11)	(0.09)	(0.11)	(0.35)	(0.39)
Grade 7	0.53	0.45	0.36	0.01	−0.04
	(0.11)	(0.10)	(0.09)	(0.37)	(0.41)
Grade 8	0.61	0.47	0.39	0.12	0.08
	(0.22)	(0.11)	(0.08)	(0.44)	(0.48)

Note: Standard errors, generated via a Bayesian bootstrap procedure, are shown in parentheses. Optimal Assignment keeps the fraction of individuals treated fixed, but assigns the individuals with the highest treatment effects to the treatment. Realized Assignment is the average treatment effect of the realized assignment, which corresponds to the average treatment on the treated (ATT). Local Effect corresponds to the effect of the program on compliers at the treatment threshold (LATE). Random Assignment is the average treatment effect if treatment was assigned randomly, which corresponds to the overall average treatment on the treated (ATE). Program Expansion is the average treatment effect if treatment expanded to the individuals not currently receiving the treatment and corresponds to the average treatment on the controls (ATC).

References

Abadie, Alberto and Matias D. Cattaneo, “Econometric Methods for Program Evaluation,” *Annual Review of Economics*, 2018, *10* (1), 465–503.

Angrist, Joshua D. and Miikka Rokkanen, “Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff,” *Journal of the American Statistical Association*, 2015, *110* (512), 1331–1344.

_ **, Peter D. Hull, Parag A. Pathak, and Christopher R. Walters**, “Leveraging Lotteries for School Value-Added: Testing and Estimation,” *Quarterly Journal of Economics*, 2017, *132* (2), 871–919.

Battistin, Erich and Enrico Rettore, “Ineligibles and eligible non-participants as a double comparison group in regression discontinuity designs,” *Journal of Econometrics*, 2008, *142* (2), 715–730.

Bertanha, Marinho, “Regression discontinuity design with many threshold,” *Journal of Econometrics*, 2020, *218* (1), 216–241.

_ **and Guido W. Imbens**, “External Validity in Fuzzy Regression Discontinuity Designs,” *Journal of Business & Economic Statistics*, 2020, *38* (3), 593–612.

Branson, Zach, Maxime Rischard, Luke Bornn, and Luke Miratrix, “A Nonparametric Bayesian Methodology for Regression Discontinuity Designs,” *Journal of Statistical Planning and Inference*, 2019, *202* (14-30).

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall, “Beyond LATE with a Discrete Instrument,” *Journal of Political Economy*, 2017, *125* (4), 985–1039.

Cattaneo, Matias D. and Rocio Titiunik, “Regression Discontinuity Designs,” *Annual Review of Economics*, 2022, *14*, 821–851.

_ **, Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare**, “Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs,” *Journal of the American Statistical Association*, 2021, *116* (536), 1941–1952.

Cerulli, Giovanni, Yingying Dong, Arthur Lewbel, and Alexander Poulsen, "Testing Stability of Regression Discontinuity Models," in Matias D. Cattaneo and Juan Carlos Escanciano, eds., *Regression Discontinuity Designs: Theory and Applications*, Vol. 38 2017, pp. 317–339.

Chetty, Raj and Nathaniel Hendren, "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates," *Quarterly Journal of Economics*, 2018, *144* (3), 1163–1228.

Chib, Siddhartha and Liana Jacobi, "Bayesian Fuzzy Regression Discontinuity Analysis and Returns to Compulsory Schooling," *Journal of Applied Econometrics*, 2016, *31* (1026-1047).

_ , Edward Greenber, and Anna Simoni, "Nonparametric Bayes Analysis of the Sharp and Fuzzy Regression Discontinuity Designs," *Econometric Theory*, 2023, *39* (3), 481–533.

Cook, Thomas D., ""Waiting for Life to Arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics," *Journal of Econometrics*, 2008, *142* (2), 636–654. The regression discontinuity design: Theory and applications.

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

Figlio, David and Umut Özek, "An extra year to learn English? Early grade retention and the human capital development of English learners," *Journal of Public Economics*, 2020, *186*, 104184.

Freyberger, Joachim and Matthew A. Masten, "A practical guide to compact infinite dimensional parameter spaces," *Econometric Reviews*, 2019, *38* (9), 979–1006.

Greene, Jay and Marcus Winters, "Revisiting grade retention: An evaluation of Florida's test-based promotion policy," *Education Finance and Policy*, 2007, *2* (4), 319–340.

Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw, "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 2001, *69* (1).

Heckman, James J., "Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 2010, *48* (2), 356–398.

_ **and Edward J. Vytlacil**, "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in J. J. Heckman and E. E. Leamer, eds., *Handbook of Econometrics*, Vol. 6, Elsevier, 2007, chapter 70, pp. 4779–4874.

_ **and** _ , "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treat- ment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments," in J. J. Heckman and E. E. Leamer, eds., *Handbook of Econometrics*, Vol. 6, Elsevier, 2007, chapter 71, pp. 4785–5143.

_ **and Edward Vytlacil**, "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 1999, *96* (8), 4730–4734.

_ **and** _ , "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 2005, *73* (3), 669–738.

Hwang, NaYoung and Cory Koedel, "Holding back to move forward: The effects of retention in the third grade on student outcomes," 2022.

Imbens, Guido W. and Thomas Lemieux, "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 2008, *142* (2), 615–635. The regression discontinuity design: Theory and applications.

Kline, Patrick and Christopher R. Walters, "On Heckits, LATE, and Numberican Equivalence," *Econometrica*, March 2019, *87* (2), 677–696.

Kowalski, Amanda, "Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform," *Review of Economics and Statistics*, 2023, *105* (3), 646–664.

Licalsi, Christina, Umut Özek, and David Figlio, "The uneven implementation of universal school policies: Maternal education and Florida's mandatory grade retention," *Education Finance and Policy*, 2019, *14* (3), 383–413.

Mealli, Fabrizia and Carla Rampichini, "Evaluating the effects of university grants by using regression discontinuity designs," *Journal of the Royal Statistical Society, Series A*, 2012, *175*, 775–798.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky, "Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters," *Econometrica*, 2018, *86* (5), 1589–1619.

Mulhern, Christine, Isaac M. Opper, Fatih Unlu, Brian Phillips, and Julie Edmunds, "Dual Method of Dual Enrollment: Combining empirical approaches to estimate the impacts of taking college courses in high school on educational attainment," 2023.

Mumma, Kirsten and Marcus Winters, "The effect of retention under Mississippi's test-based promotion policy," 2023.

Newey, Whitney K. and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in R.F. Engle and D.L. McFadden, eds., *Handbook of Econometrics, Volume IV*, Elsevier Science, 1994.

Opper, Isaac M., "From LATE to ATE: A Bayesian Approach," 2023.

Özek, Umut, "Hold back to move forward? Early grade retention and student misbehavior," *Education Finance and Policy*, 2015, *10* (3), 350–377.

Rasmussen, Carl Edward and Chrisopher K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006.

Rokkanen, Miikka, "Exam schools, ability, and the effects of affirmative action: latent factor extrapolation in the regression discontinuity design," 2015.

Schwerdt, Guido, Martin West, and Marcus Winters, "The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida," *Journal of Public Economics*, 2017, *152*, 154–169.

Vytlacil, Edward, "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 2002, *71* (1), 331–341.

Wing, Coady and Thomas D. Cook, "Strengthening the Regression Discontinuity Design Using Additional Design Elements: A Within-Study Comparison," *Journal of Policy Analysis and Management*, 2013, *32* (4), 853–877.

Winters, Marcus and Jay Greene, "The medium-run effects of Florida's test-based promotion policy," *Education Finance and Policy*, 2012, *7* (3), 305–330.

Wood, Simon N., *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC, 2017.

A Proofs

Proposition 1. *Under assumptions 1-5, the Global RDD as defined in Definition can be implemented using the observed conditional moments and is well-defined.*

Proof. The fact that the Global RDD can be estimated using the observed data is clear, so we focus the proof on showing that it is well-defined, i.e., that there is a single choice of $(\hat{y}_0, \hat{y}_1) \in \mathcal{Y}^{GRDD}$ such that $\hat{y}_k(\nu(Z), Z) = \mathbb{E}[T_i = k, Z_i = Z]$ for all $Z \neq Z^*$ and $k \in \{0, 1\}$. We will focus on the case where $k = 0$, but the proof for the case of $k = 1$ is identical.

To do so, we use the fact that even though we do not observe the conditional moments at Z^*, any possible $\hat{y}_0 \in \mathcal{Y}_0^{GRDD}$ that satisfies $\hat{y}_0(\nu(Z), Z) = \mathbb{E}[T_i = 0, Z_i = Z]$ for all $Z \neq Z^*$ needs satisfy the restriction that $\hat{y}_0(p_l, Z^*) = y_0^*(p_l, Z^*)$ and $\hat{y}_0(p_h, Z^*) = y_0^*(p_h, Z^*)$. If not, it would be impossible to choose a continuous function $\gamma(Z)$ such that both: $\hat{y}_0(\nu(Z^* + \epsilon), Z^* + \epsilon) = \mathbb{E}[T_i = 0, Z_i = Z^* + \epsilon]$ and $\hat{y}_0(\nu(Z^* - \epsilon), Z^* - \epsilon) = \mathbb{E}[T_i = 0, Z_i = Z^* - \epsilon]$ for a small ϵ.

We can then use that there is a single choice of β_0^* that goes through both $y_0^*(p_h, Z^*)$ and $y_0^*(p_l, Z^*)$. Since Z^* is the only point where we observe multiple values of $y_0^*(\eta, Z)$, we can then set $\gamma^*(Z) = \mathbb{E}[Y_i|\nu(Z), Z, T_i = 0] - \beta_0^* \nu(Z)$ to ensure that $\hat{y}_k(\nu(Z), Z) = \mathbb{E}[T_i = 0, Z_i = Z]$ for all $Z \neq Z^*$. Finally, from the assumption that $\nu(Z)$ and $\mu^*(\eta, Z)$ are both continuous functions, it follows that $\gamma^*(Z)$ is a continuous function and hence $\beta_0^* \eta + \gamma^*(Z) \in \mathcal{Y}_0^{GRDD}$.

□

Remark 1. *Let τ^*_{GRDD} be the estimate generated by the Global Regression Discontinuity Design, as defined in Section II.B, and $\tau^*_{obs}(Z)$ be the estimate generated from the traditional observational study, as defined in Equation (7). We then have:*

$$\tau^*_{GRDD}(Z) = \tau^*_{obs}(Z) - b \tag{11}$$

where b is a measure of the bias in the observational estimates. Specifically, defining β_0^* and β_1^* as in Equation (9) and (10), we have:

$$\begin{aligned} b &= \beta_0^* \cdot \nu(Z) + \beta_1^* \cdot \big(1 - \nu(Z)\big) && (12) \\ &= \xi_h \cdot \big(\tau^*_{RDD} - \tau^*_{obs}(Z_h^*)\big) + \xi_l \cdot \big(\tau^*_{RDD} - \tau^*_{obs}(Z_l^*)\big) && (13) \end{aligned}$$

where $\xi_k \in \mathbb{R}$ *is a function of* p_h, p_l *and* $\nu(Z)$*,* $\tau^*_{obs}(Z^*_h) = \lim_{Z \downarrow Z^*} \tau^*_{obs}(Z)$*, and* $\tau^*_{obs}(Z^*_l) = \lim_{Z \uparrow Z^*} \tau^*_{obs}(Z)$*.*

Proof. We start by noting that in the Global RDD we restrict the functional form of the estimated moments to be of the form: $\hat{y}_0(\nu(Z), Z) = \beta^*_0 \nu(Z) + \gamma(Z)$ and $\hat{y}_1(\nu(Z), Z) = \beta^*_1 \nu(Z) + \delta(Z)$. Thus, we get that:

$$\begin{aligned}\tau^*_{GRDD}(Z) &= \hat{y}_1(1, Z) - \hat{y}_0(0, Z) \\ &= \Big(\hat{y}_1(1, Z) - \hat{y}_1(\nu(Z), Z)\Big) + \Big(\hat{y}_1(\nu(Z), Z) - \hat{y}_0(\nu(Z), Z)\Big) + \Big(\hat{y}_0(\nu(Z), Z) - \hat{y}_0(0, Z)\Big) \\ &= \big(1 - \nu(Z)\big) \cdot \beta^*_1 + \tau^*_{obs}(Z) + \nu(Z) \cdot \beta^*_0\end{aligned}$$

which gives the first expression for the bias.

To connect the bias measure to the traditional RDD, we start by rearranging the traditional RD estimate to write that it can be expressed as either:

$$\tau^*_{RDD} = \tau^*_{obs}(Z^*_l) + \frac{y^*_0(p_h, Z^*) - y^*_0(p_l, Z^*)}{p_h - p_l} \cdot (1 - p_h) + \frac{y^*_1(p_h, Z^*) - y^*_1(p_l, Z^*)}{p_h - p_l} \cdot p_h \quad (26)$$

or

$$\tau^*_{RDD} = \tau^*_{obs}(Z^*_h) + \frac{y^*_0(p_h, Z^*) - y^*_0(p_l, Z^*)}{p_h - p_l} \cdot (1 - p_l) + \frac{y^*_1(p_h, Z^*) - y^*_1(p_l, Z^*)}{p_h - p_l} \cdot p_l \quad (27)$$

where y^* are the true conditional moments at the specified points.

As before, we have that β^*_0 and β^*_1 are defined such that:

$$\beta^*_0 = \frac{y^*_0(p_h, Z^*) - y^*_0(p_l, Z^*)}{p_h - p_l} \qquad \text{and} \qquad \beta^*_1 = \frac{y^*_1(p_h, Z^*) - y^*_1(p_l, Z^*)}{p_h - p_l} \quad (28)$$

Thus, we get that we can re-write Equations (26) and (27) as the set of linear equations:

$$\begin{bmatrix} 1 - p_h & p_h \\ 1 - p_l & p_l \end{bmatrix} \begin{bmatrix} \beta^*_0 \\ \beta^*_1 \end{bmatrix} = \begin{bmatrix} \tau^*_{RDD} - \tau^*_{obs}(Z^*_l) \\ \tau^*_{RDD} - \tau^*_{obs}(Z^*_h) \end{bmatrix} \quad (29)$$

From this it is clear that we can write $\beta_{\substack{*\\0}}$ as a linear combination of $\tau^*_{RDD} - \tau^*_{obs}(Z^*_l)$ and $\tau^*_{RDD} - \tau^*_{obs}(Z^*_h)$, with the weights depending on p_h and p_l, and that the same is true (with different weights) for $\beta_{\substack{*\\1}}$. Plugging that into the expression that:

$$\tau^*_{GRDD}(Z) = \big(1 - \nu(Z)\big) \cdot \beta^*_1 + \tau^*_{obs}(Z) + \nu(Z) \cdot \beta^*_0 \quad (30)$$

we get that we can also write:

$$\tau^*_{GRDD}(Z) = \xi_h \cdot \big(\tau^*_{RDD} - \tau^*_{obs}(Z^*_h)\big) + \xi_l \cdot \big(\tau^*_{RDD} - \tau^*_{obs}(Z^*_l)\big) \tag{31}$$

where $\xi_k \in \mathbb{R}$ is a function of p_h, p_l and $\nu(Z)$.

□

Proposition 2. *Let τ^* denote the true MTE function. Then the estimated effect on the set of compliers at the Z^* converges to the true effect on that set, i.e.:*

$$\frac{1}{p_h - p_l} \int_{p_l}^{p_h} \tau^*_{GRDD}(\eta, Z^*) d\eta = \frac{1}{p_h - p_l} \int_{p_l}^{p_h} \tau^*(\eta, Z^*) d\eta \tag{14}$$

Proof. We start by noting that:

$$\int_{p_l}^{p_h} \tau^*(\eta, Z^*) d\eta = \Big(p_h y^*_1(p_h, Z^*) - p_l y^*_1(p_l, Z^*)\Big) - \Big((1-p_l) y^*_0(p_l, Z^*) - (1-p_h) y^*_0(p_l, Z^*)\Big)$$

and we can similarly write

$$\int_{p_l}^{p_h} \tau^*_{GRDD}(\eta, Z^*) d\eta = \Big(p_h \hat{y}_1(p_h, Z^*) - p_l \hat{y}_1(p_l, Z^*)\Big) - \Big((1-p_l) \hat{y}_0(p_l, Z^*) - (1-p_h) \hat{y}_0(p_l, Z^*)\Big)$$

where $\hat{y}$ corresponds to the estimated moments in the Global Regression Discontinuity Design.

From the proof of Proposition 1, however, it follows that $\hat{y}(\eta, Z)$ equals $y^*(\eta, Z)$ at every observed point, i.e., at every point $(\nu(Z), Z)$, and at both (p_h, Z^*) and (p_l, Z^*). The theorem thus follows. □

Proposition 3. *Let y^* be the true conditional moments and suppose that $y^* \in \mathcal{Y}^{GRDD}$. Then the estimated MTE function converges to the true MTE function, i.e., $\tau^*_{GRDD} = \tau^*$.*

Proof. As shown in Proposition 1, we have that the $\hat{y}$ resulting from the Global Regression Discontinuity Design are the only $y \in \mathcal{Y}^{GRDD}$ that matches y^* at the observed moments. If $y^* \in \mathcal{Y}^{GRDD}$, then $\hat{y}$ must match y^* at every moment and so $\tau^*_{GRDD} = \tau^*$. □

Proposition 4. *The assumptions under which the Global RDD converges to the true treatment effect function are weaker than those under which the observational study* *does. Similarly, the conditions under which the Global RDD converges to the true average treatment effect on a complier population are weaker than the conditions under which the Angrist and Rokkanen (2015) estimator does.*

Proof. We start by formally defining a version of the Global RDD that includes a set of covariates, which we denote as X_i for individual i.[17] To extend the model, we will let $\nu(Z, X) = Pr(T_i = 1|Z_i = Z, X_i = X)$, as well as $p_l(X) = \lim_{Z \uparrow Z^*} \nu(Z, X)$ and $p_h(X) = \lim_{Z \downarrow Z^*} \nu(Z, X)$.[18] Finally, we will again set $y_k(\eta, Z, X) \equiv \mathbb{E}[Y_i|\eta_i = \eta, Z_i = Z, X_i = X]$.

It can then be defined as:

$$\tau^*_{GRDD}(\eta, Z, X) = \tau^*_{obs}(Z, X) - \beta^*_{0,X} \cdot \nu(Z, X) - \beta^*_{1,X} \cdot \big(1 - \nu(Z, X)\big) \tag{32}$$

where $\tau^*_{obs}(Z, X) = \mathbb{E}[Y_i|T_i = 1, Z_i = Z, X_i = X] - \mathbb{E}[Y_i|T_i = 1, Z_i = Z, X_i = X]$, $\beta^*_{0,X} = \frac{y_0(p_h(X),Z^*,X)-y_0(p_l(X),Z^*,X)}{p_h(X)-p_l(X)}$ and $\beta^*_{1,X} = \frac{y_1(p_h(X),Z^*,X)-y_1(p_l(X),Z^*,X)}{p_h(X)-p_l(X)}$.

From here, the proof is straightforward. First, it follows that $\tau^*_{obs}(Z, X) = \tau^*(Z, X)$ if and only if $y_k(\eta, Z, X)$ does not depend on η for $k \in \{0, 1\}$. If that is the case, then $\beta^*_0(X) = \beta^*_1(X) = 0$ and so $\tau^*_{GRDD}(\eta, Z, X) = \tau^*_{obs}(Z, X) = \tau^*(\eta, Z, X)$. It also clear that there exist formulations of $y_k(\eta, Z, X)$, namely those that are additively separable in η and (Z, X) and linear in η, under which $\tau^*_{GRDD}(\eta, Z, X) = \tau^*(\eta, Z, X)$ but where $\tau^*_{obs}(Z, X) \neq \tau^*(Z, X)$. Thus, the assumption required for the Global RDD to converge to the true conditional average treatment effect function is strictly weaker than the assumption required for the observational study to do so.

Next, we turn to the assumptions of Angrist and Rokkanen (2015). In our notation, their *Generalized conditional independence assumption (GCIA)* can be written as the assumption that:

$$y_k(\eta, Z, X) = y_k(\eta, X) \tag{33}$$

$$\nu(Z, X) = p_h(X) \cdot \mathbf{1}(Z_i > Z^*) + p_l(X) \cdot \mathbf{1}(Z_i < Z^*) \tag{34}$$

and the *Conditional first stage* being that $p_h(X) \neq p_l(X)$ for all X.[19] Under this

[17] Roughly speaking, this formulation simply involves separately estimating the Global RDD as specified in Section II separately for each set of potential covariates. In practice, estimation of such a model is likely be infeasible and so we implement a different approach to "control" for covariates in Section IV.

[18] We extend Assumption 4 to be that $p_h(X) \neq p_l(X)$ for all X.

[19] As shown in Vytlacil (2002), the monotonicity assumption is implicit in the generalized Roy

assumption, it follows that:

$$\frac{1}{p_h(X) - p_l(X)} \int_{p_l(X)}^{p_h(X)} \tau^*_{GRDD}(\nu, X) d\nu = \frac{1}{p_h(X) - p_l(X)} \int_{p_l(X)}^{p_h(X)} \tau^*(\nu, X) d\nu \quad (35)$$

for any X, with the proof being identical to the one for Proposition 2. Again, it is also clear that there exist formulations of $y_k(\eta, Z, X)$ such that τ^*_{GRDD} converges to the true treatment effect function, but the Angrist and Rokkanen estimator does not. Thus, the assumption required for the Global RDD to converge to the true local average treatment effect are strictly weaker than the assumptions required for the Angrist and Rokkanen estimator to do so. □

Proposition 5. *For any choice of C_0 and C_1 and any point $\tilde{Z} \neq Z^*$, we get that:*

$$\tau^*_{GRDD}(\eta, Z) = \mathbb{E}\big[\tau^*(\eta, Z) | \mathcal{D}(\{\tilde{Z}, Z^*\})\big] \quad (18)$$

for every η and $\in \{\tilde{Z}, Z^*\}$.

Proof. The proof is straightforward, but requires some additional notation. First, we will let $h(\eta, Z)$ be a 3×1 vector equal to $[1, \eta, Z]$ and θ to be 3×1 coefficient vector equal to $[\alpha_k, \beta_k, \gamma_k]'$. Thus, the linear portion of the modified GP can be written as $h(\eta, Z)'\theta$. We also let H denote a matrix constructed by stacking the values of $h(\eta, Z)'$ of all the observed data and Y denote a vector of the observed outcomes.

Next, we introduce a succinct way to denote the covariance terms of the rest of the GP. If we use (as in the main paper) $\mathcal{D}$ to denote the observed data and let$N$ be the number of observed data points, we then let $C_k\big((\eta, Z), \mathcal{D}\big)$ be a $N \times a$ vector where the i^{th} row is equal to $C_k\big((\eta, Z), (\eta_i, Z_i)\big)$. We similarly let $C_k\big(\mathcal{D}\big)$ be a $N \times N$ matrix where the (i, j) value of the matrix is equal to $C_k\big((\eta_i, Z_i), (\eta_j, Z_j)\big)$.

Given this notation, we can write the mean posterior as:

$$\mathbb{E}\Big[y_k(\eta, Z) \big| \mathcal{D}\big(\{\tilde{Z}, Z^*\}\big)\Big] = h(\eta, Z)'\hat{\theta} + C_k\big((\eta, Z), \mathcal{D}\big)' C_k\big(\mathcal{D}\big)^{-1} r \quad (36)$$

where $\hat{\theta} = \big(H'C_k(\mathcal{D})^{-1}H\big)^{-1} H'C_k(\mathcal{D})^{-1}Y$ and $r = Y - H'\hat{\theta}$.

Note that since we condition only on $\mathcal{D}\big(\{\tilde{Z}, Z^*\}\big)$, we only consider the case in which we observe three data points: $(\nu(\tilde{Z}), \tilde{Z})$, (p_l, Z^*), and (p_h, Z^*). The linear

model introduced in Section II.A.

model can therefore perfectly explain the observed outcomes, i.e., $r = 0$, which implies that: $\mathbb{E}\Big[y_k(\eta, Z)\big|\mathcal{D}\big(\{\tilde{Z}, Z^*\}\big)\Big] = h(\eta, Z)'\hat{\theta}$.

As outlined in Proposition 1, we also know that the only way to perfectly fix the observed outcomes is to have the slope on the η term equal $\beta_k^* = \frac{y_k(p_h, Z^*) - y_k(p_l, Z^*)}{p_h - p_l}$ and so we get that:

$$\mathbb{E}\Big[y_k(\eta, Z)\big|\mathcal{D}\big(\{\tilde{Z}, Z^*\}\big)\Big] = (\eta - \nu(Z))\beta_k^* + y_k\big(\nu(Z), Z\big) \tag{37}$$

for $Z \in \{\tilde{Z}, Z^*\}$, which gives the equivalent formulation as $\hat{y}_k(\eta, Z)$ as generated by the Global RDD. From the fact that T is linear, we therefore get that $\mathbb{E}[\tau^*(\eta, Z)|\mathcal{D}\big(\{\tilde{Z}, Z^*\}\big)] = \tau^*_{GRDD}(\eta, Z)$ for all η and $Z \in \{\tilde{Z}, Z^*\}$. □

Proposition 6. *Define to $b^*_{obs}(\eta, Z)$ be the bias in the observational study, i.e.,*

$$b^*_{obs}(\eta, Z) = \tau^*_{obs}(Z) - \tau^*(\eta, Z) \tag{19}$$

where $\tau^*_{obs}(Z)$ is defined in Equation (7). Then:

$$\tau^*_{GRDD}(\eta, Z) = \tau^*_{obs}(Z) - \mathbb{E}\big[b^*_{obs}(\eta, Z)\big|\mathcal{D}(\{Z, Z^*\})\big] \tag{20}$$

for any (η, Z) and any choice of C_0 and C_1.

Proof. This follows immediately from the fact that $\mathbb{E}[\tau^*(\eta, Z)|\mathcal{D}\big(\{\tilde{Z}, Z^*\}\big)] = \tau^*_{GRDD}(\eta, Z)$ – as shown in the Proposition 5 – and the definition of $b^*_{obs}(\eta, Z)$.

□

Proposition 7. *Suppose that $C_k((\eta, Z), (\eta', Z')) = C_{k,\eta}(\eta, \eta') + C_{k,Z}(Z, Z')$ for both $k \in \{0, 1\}$ and that:*

$$\nu(Z) = \begin{cases} p_l \text{ if } Z < Z^* \\ p_h \text{ if } Z > Z^* \end{cases} \tag{21}$$

Then for any choice of $C_{k,\eta}$ and $C_{k,Z}$, we get that:

$$\tau^*_{GRDD}(\eta, Z) = \mathbb{E}\big[\tau^*(\eta, Z)|\mathcal{D}(\mathbf{Z})\big] \tag{22}$$

for every (η, Z).

Proof. Using the notation introduced in the above proof, we start by re-casting mean posterior of the Gaussian process as:

$$\mathbb{E}\big[y_k(\eta, Z)\big|\mathcal{D}(\mathbf{Z})\big] = h(\eta, Z)'\hat{\theta} + \hat{\tilde{y}}(\eta, Z) \tag{38}$$

where the parameters $\hat{\theta}$ and $\hat{\tilde{y}}$ are the solutions to:

$$\arg\min_{\theta, y} ||y|| \text{ s.t. } h(\nu(Z), Z)'\theta + y(\nu(z), Z) = y_k(\nu(Z), Z)\ \forall Z \tag{39}$$

and the norm $||y||$ depends on the chosen kernel C_k. Next, from the assumption that $C_k = C_{k,\eta} + C_{k,Z}$ we know that $\tilde{y}(\eta, Z) = f(\eta) + g(Z)$ and $||y|| = ||f|| + ||g||$.

We then use the assumption that $\nu(Z)$ is a step function to get that $g(Z)$ is is completely determined by $\hat{\alpha}_k$, $\hat{\delta}_k$, and the constraint that $h(\nu(Z), Z)'\theta + y(\nu(z), Z) = y_k(\nu(Z), Z)$ for all Z. Importantly, this means that we choose β_k and f without consideration of α_k, δ_k and g, i.e., we can re-write Equation 39 as:

$$\arg\min_{\beta_k, f} ||f|| \text{ s.t. } \beta_k\nu(Z) + f(\nu(Z)) = y_k(\nu(Z), Z) - \big(\alpha_k + \gamma_k Z + g(Z)\big)\ \forall Z \tag{40}$$

Finally, as outlined in Proposition 1, there is a unique β_k^*, equal to $\frac{y_k(p_h, Z^*) - y_k(p_l, Z^*)}{p_h - p_l}$, that allows $||f|| = 0$ and hence satisfies the minimization. Thus, we get that:

$$\mathbb{E}\big[y_k(\eta, Z)\big|\mathcal{D}(\mathbf{Z})\big] = (\eta - \nu(Z))\beta_k^* + y_k\big(\nu(Z), Z\big) \tag{41}$$

which gives the equivalent formulation as $\hat{y}_k(\eta, Z)$ as generated by the Global RDD. From the fact that T is linear, we therefore get that $\mathbb{E}[\tau^*(\eta, Z)|\mathcal{D}] = \tau^*_{GRDD}(\eta, Z)$ for all η and Z.

□

Proposition 8. *Let $\hat{\tau}_{GRDD}$ be the estimate generated by the Global Regression Discontinuity Design, as defined above, and $\hat{\tau}^*_{GRDD}$ be the MTE function defined in Definition 1. Then given Assumptions 1-4 and 6-9 we get that:*

$$\hat{\tau}_{GRDD} \xrightarrow{p} \tau^*_{GRDD} \tag{25}$$

Proof. For notation, we start using y to denote the parameters, i.e., $y = (\beta_0, \beta_1, \gamma, \delta)$, and then define:

$$Q_n(y) = \frac{1}{N}\sum_{\forall i}\left(Y_i - \Big(\beta_0 \cdot \hat{\nu}(Z_i) + \gamma(Z_i) + T_i \cdot \big((\beta_1 - \beta_0)\cdot \hat{\nu}(Z_i) + \delta(Z_i)\big) + \xi X_i\Big)\right)^2 \\ + \frac{1}{N}\lambda_\gamma \int \Big(\gamma''(Z)\Big)^2 dz + \frac{1}{N}\lambda_\delta \int \Big(\delta''(Z)\Big)^2 dZ$$

and

$$Q_0(y) = \mathbb{E}\left[\Big(Y_i - \beta_0\nu(Z) + \gamma(Z) + T_i \cdot \big((\beta_1 - \beta_0)\cdot \nu(Z) + \delta(Z)\big)\Big)^2\right] \tag{42}$$

We then show that four assumptions in Theorem 2.1 of Newey and McFadden (1994) hold, i.e., that: (i) Q_0 has a unique minimum; (ii) the parameter space is compact; (iii) Q_0 is continuous; and (iv) $sup_{y\in\mathcal{Y}^{GRDD}}|\hat{Q}_n(y) - Q_0(y)| \xrightarrow{p} 0$. Since the four assumptions hold, we can then conclude that $\hat{y} \xrightarrow{p} \hat{y}^*$ for the unique $\hat{y}^* \in \mathcal{Y}^{GRDD}$ that minimizes $Q_0(y)$.

The proofs that condition (iii) holds is straightforward to show and condition (ii) follows from Assumption 9. Furthermore, from the fact that $\hat{\nu} \to \nu$ – which in turn can be proved using the same approach as in this proof – and the law of large numbers, we get that condition (iv) holds.

The only condition whose proof is unique to the this context is (i), i.e., that Q_0 has a unique minimum when we restrict the possible functions to the set $\mathcal{Y}_{GRDD}$. This follows from Proposition 1, as does the fact that the $\hat{y}^* \in \mathcal{Y}^{GRDD}$ that minimizes $Q_0(y)$ corresponds to the values $(\hat{y}_0, \hat{y}_1) \in \mathcal{Y}^{GRDD}$ such that $\hat{y}_k(\nu(Z), Z) = \mathbb{E}[Y_i|T_i = k, Z_i = Z]$ for all Z.

Finally, since T is continuous, we can conclude that $\hat{\tau}_{GRDD} = T(\hat{y}) \xrightarrow{p} T(\hat{y}^*) \equiv \tau^*_{GRDD}$. □

Proposition 9. *For any choice of C_0 and C_1, the expected loss is lower for the Global Regression Discontinuity Design than for a traditional observational study.*

Proof. From the law of iterated expectations, we can then re-write the expected loss for the observational study as:

$$\mathcal{L} = \mathbb{E}\left\{\mathbb{E}\left[\Big(\tau^*(\eta, Z) - \tau^*_{obs}(\eta, Z)\Big)^2 \middle| \mathcal{D}\Big(\{Z^*, Z\}\Big)\right]\right\} \tag{43}$$

for any $Z \neq Z^*$. We can then focus on the inside expectation and show that:

$$\mathbb{E}\left[\Big(\tau^*(\eta,Z) - \tau^*_{obs}(\eta,Z)\Big)^2 \middle| \mathcal{D}\big(\{Z^*,Z\}\big)\right] \geq \mathbb{E}\left[\Big(\tau^*(\eta,Z) - \tau^*_{GRDD}(\eta,Z)\Big)^2 \middle| \mathcal{D}\big(\{Z^*,Z\}\big)\right] \tag{44}$$

for all $\mathcal{D}\big(\{Z^*,Z\}\big)$. This follows directly from Proposition 5. Since the expected loss conditional on $\mathcal{D}\big(\{Z^*,Z\}\big)$ is lower for the Global RDD than the traditional observational study regardless of $\mathcal{D}\big(\{Z^*,Z\}\big)$, the expected loss is also lower unconditionally. □

Proposition 10. *Decompose the modified Gaussian process as defined in Section III so that we can separate the functions into: (a) one capturing the direct effect of* η*, (b) one capturing the direct effect of* Z*, and (c) one capturing the interaction, i.e., let*

$$y_k(\eta,Z) = \tilde{y}_{k,\eta}(\eta) + \tilde{y}_{k,Z}(Z) + \tilde{y}_{k,\eta,Z}(\eta,Z) \tag{45}$$

Then holding fixed the rest of the the Gaussian process, if:

$$\mathbb{V}\Big(\big(\tilde{y}_{1,Z}(Z) - \tilde{y}_{0,Z}(Z)\big) - \big(\tilde{y}_{1,Z}(Z^*) - \tilde{y}_{0,Z}(Z^*)\big)\Big) \tag{46}$$

is large enough for $Z \neq Z'$*, then the expected loss is lower for the Global RDD than for a traditional fuzzy RDD.*

Proof. Since the Global RDD involves comparing $y_1(\nu(\eta),Z)$ to $y_0(\nu(\eta),Z)$, the expected loss of the Global RDD does not depend on the value of Equation 46. In contrast, since the traditional fuzzy RDD only uses information at the discontinuity, an increase of $\mathbb{V}\Big(\big(\tilde{y}_{1,Z}(Z) - \tilde{y}_{0,Z}(Z)\big) - \big(\tilde{y}_{1,Z}(Z^*) - \tilde{y}_{0,Z}(Z^*)\big)\Big)$ increases the expected loss of the fuzzy RDD by the same amount, holding the rest of the modified GP fixed. Thus, for a large enough value of Equation 46 the expected loss of the fuzzy RDD is larger than for the Global RDD. □

B Extensions

An advantage of using a marginal treatment effect representation of the fuzzy RDD setting is that the flexibility provides a number of possible extensions to the method presented above. While we will mostly leave these for future study, we briefly touch on some of these extensions here as guidance for those researchers who want to apply the method to one of these contexts.

B.A Multiple Discontinuities

There are many contexts in which there are multiple discontinuities and a growing body of literature investigates how best to handle these cases (e.g., Cattaneo et al. (2021),Bertanha (2020)). While our focus has been contexts where there is a single discontinuity, as specified in Assumption 4, the Global RDD can also be applied to the case where there are multiple discontinuities. We briefly discuss here how the presence of multiple discontinuities changes the results in Section II and III and then use this to illustrate how we can extend the Global RDD to handle multiple discontinuities.

The assumption that there is a single discontinuity was important for the previous results because it ensured that the definition of τ^*_{GRDD} is well-defined, i.e., that there exists a unique function $\hat{y}_k \in \mathcal{Y}_k^{GRDD}$ such that $\hat{y}_k(\nu(Z), Z) = \mathbb{E}[Y_i|T_i = k, Z_i = Z]$ for all $Z \neq Z^*$ and $k \in \{0, 1\}$. In other words, the issue with there being multiple discontinuities is that it is no longer guaranteed that there exists an additively separable and linear in η conditional moment function that can match all of the observed moments. In response to this issue we have essentially two choices. We can either: (a) relax the assumption that they are additively separable and linear in η in a way that ensures there still exists a single function $\hat{y}_k \in \mathcal{Y}_k^{GRDD}$ that matches every observed moment or (b) use the multiple discontinuities to improve the precision with which we estimate $\hat{\beta}_0$ and $\hat{\beta}_1$.

As an example of the first approach, with multiple discontinuities we can still restrict the functions $\hat{y}_k(\eta, Z)$ to be additively separable but allow the η term to be a higher-order polynomial, with the order depending on how many discontinuities there are (and how the probability of treatment changes at these points). In this case, how much allowable flexibility in the η term we could allow due to the multiple discontinuities is more or less identical to the analysis in Brinch et al. (2017). Alternatively, we could still restrict that functions $\hat{y}_k(\eta, Z)$ to be linear in η (for any Z), but allow the slope to depend on Z. For example, we could specify that $y^*_k(\eta, Z) = \delta(Z) + \beta(Z)\eta$, where $\beta(Z)$ is some polynomial of Z; again, the flexibility in our specification of $\beta(Z)$, e.g., whether we allow it to be a constant, linear function of Z, or higher-order polynomial, would depend on how many discontinuities there are.

An alternative is to keep the estimation approach specified in Section IV even in the presence of multiple discontinuities. To be clear, doing so invalidates many of the results that rely on the Bayesian model presented in Section III. However, the multiple discontinuities would lead to a more precise estimation of an "average slope" and we believe this will likely preferable in most contexts. We also note that, if one continuous to restrict the estimated moments to be additively separable and linear in η, the multiple discontinuities lead to the model being over-identified. This, in turn, provides the ability to empirically test the null hypothesis that the true model is additively separable and linear in η. We next briefly discuss more generally how the MTE representation of the fuzzy RD provides the ability to test a range of more restricted models.

B.B Testing Restricted Models

The main issue we have discussed in this paper is that the functions $y_k(\eta, Z)$ are not identified in the fuzzy RD context without significant additional restrictions. We have therefore proposed a particular restriction and then spent Section II discussing how this transforms the observed moments into the resulting estimates and III discussing how to motivate this restriction. One downside of our choice of restrictions is that because it perfectly explains the observed moments, there is no way to test empirically whether this restriction is plausible or not.

We now turn briefly to an alternative use of the MTE representation, which is that it suggests natural ways to test for potentially interesting null hypotheses.[20] Roughly speaking, by further restricting the set of plausible conditional moments we can get an over-identified model, and therefore can conduct empirical tests of the restricted model. We now discuss four null hypotheses, which we find are often of interest to researchers and which we return as part of the accompanying R package.

For the first restricted model, we consider testing the null hypothesis that there is no endogenous selection into treatment. In the MTE model, this is equivalent to testing the null hypothesis that $\hat{\beta}_0 = \hat{\beta}_1 = 0$. While we feel like it is worthwhile to mention this test and to include it in the output generated by the R package, we want to highlight that this is the identical test as proposed in Bertanha and Imbens (2020) and so encourage those particularly interested in testing this null to see Bertanha and

[20] Many thanks to the participants at the 2023 AEFP for their thoughtful comments on a (very!) early version of this paper, which in turn motivated this section.

Imbens (2020) for more details on the test.[21] Note also that, at least asymptotically, the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are determined only using information at the discontinuity, e.g., see Equations (9) and (10).

For the second restricted model, we relax the restriction that there is no endogenous selection and instead test the null hypothesis that there is linear endogenous selection and constant treatment effects. That is, we test the joint null hypothesis that $\tau^*(\eta, Z) = \tau$ and $\mu^*(\eta, Z) = \alpha\eta + \tilde{\gamma}(Z)$ for some $\alpha \in \mathbb{R}$ and $\tilde{\gamma} : \mathbf{Z} \to \mathbb{R}$. It is easy to show that this restriction corresponds the restriction that $\hat{\beta}_1 = \hat{\beta}_0$ and $\hat{\delta}(Z) = 0$ for all Z and so therefore can be easily tested using the results from Step 2 if the Global Regression Discontinuity Design outlined in Section IV. Unlike the previous test we discussed, this null uses information both at the discontinuity – to test that $\hat{\beta}_1 = \hat{\beta}_0$ – and away from the discontinuity – to test that $\hat{\delta}(Z) = 0$). However, it is also worth emphasizing that rejecting the null hypothesis does not allow researchers to know whether that is due to heterogeneous treatment effects (which we believe is the null hypothesis they are generally interested in) or to non-linear endogenous selection (which is an additional restriction needed to over-identify the model). Still, we believe that rejecting the null (or failing the reject the null) is informative and so also include the results of this null hypothesis in the R output. We also include a test of the related null hypothesis that there is linear endogenous selection and no treatment effect for anyone.

Finally, the last two tests allow for some treatment effect heterogeneity, but restrict the form it takes. For one of the hypothesis tests, we allow for treatment effect heterogeneity in the unobserved propensity to enroll but not in Z, i.e., to relax the previous restriction to be that $\tau^*(\eta, Z) = \alpha_1\eta$ for some $\alpha_1 \in \mathbb{R}$. For the second, we allow for treatment effect heterogeneity in the running variable but not in η, i.e., we relax the previous restriction to be that $\tau^*(\eta, Z) = \delta(Z)$ for some continuous function δ. Note that, as in the previous test, both of these test the join null that the treatment effect heterogeneity is restricted and that there is linear endogenous selection.

[21] This is true asymptotically; in practice, the different ways in which we estimate the observed conditional moments mean that the results may differ slightly in a finite-samples. If all you are interested in is testing this null, we suggest you use the package put together by Bertanha and Imbens (2020).

B.C One-Sided Non-Compliance

A common source of fuzzy regression discontinuity designs are cases where there is an eligibility threshold, but where not everyone who meets the threshold actually enrolls in the treatment. In many of these cases a fraction of those individuals who do not meet the eligibility threshold end up enrolling anyway, in which case we can apply the Global Regression Discontinuity Design outlined above. In others, however, only those who meet the threshold end up enrolling and so there is only one-sided non-compliance. This invalidates the part of Assumption 4 where we make the assumption that $\nu(Z) \in (0, 1)$ for all Z, which seemingly suggests that we can not apply the method to these contexts.

To discuss how we might be able to extend the method to these cases, we highlight the two reasons why the overlap assumption is important. The first, and most straightforward, is that without knowledge of how $\mathbb{E}[Y_i|T_i = 1, Z_i = Z]$ varies according to Z it would be impossible to identify $\delta(Z)$. The second is more specific to the regression discontinuity context; without two-sided imperfect compliance, we only observe three conditional moments at the discontinuity rather than four, and hence are unable to identify both β_1 and β_0.

That said, we can use the logic outlined above to propose an approach that can be applied to fuzzy RDs, even in the case of one-sided noncompliance. In particular, a natural approach in these cases is to further restrict ourselves to the case in which $\hat{\beta}_1 = \hat{\beta}_0$, i.e., restrict to consider the case where there is no treatment effect heterogeneity in η, and only focus on the values of the running variable where this is overlap. This would allow us to identify, for example, the average treatment effect on the treated (ATT) and can still be motivated as the "best guess" of the treatment effects under the Bayesian model outlined in Section III. The key difference is that we should implicitly be less confident in these results since there is less data to condition on. Similarly, it can be viewed as requiring weaker assumptions to identify the ATT than the an observational study (or Angrist and Rokkanen (2015)), although it requires stronger assumptions than if there is two-sided non-compliance.

B.D Improving Precision at the Cost of Bias

As discussed in Section II, the proposed estimator can be thought of as a linear combination of: (a) a potentially biased observational study using data away from the discontinuity and (b) a consistent (but local) traditional regression discontinuity design using data at the discontinuity. Note that in this framing, it appears similar to other designs that combining observational data with quasi-experimental data; however, the motivation for these usually stem from the fact that the quasi-experimental estimates are less precise than the observational estimates and so the researcher aims to improve mean-square error of the estimates by reducing the variance at the expense of moderate increases in the bias (e.g., Angrist et al. (2017) and Chetty and Hendren (2018)). Here, in contrast, the weights reflect the fact that even without statistical uncertainty, neither the observational estimate nor the RD estimate is perfect; the observational estimates is biased due to selection bias and RD estimate is local to the complier population at the cutoff.

In practice, of course, it is often the case that the RD estimate is not only local, but imprecise. If one is concerned about the imprecision of the fuzzy RD estimates, it makes sense to further reduce the weight the on the RD estimates, thereby moving the resulting estimates toward the observational estimates. While the intuition is straightforward, it is not immediately obvious how should should do so given the complex nature of ξ_h and ξ_l in Remark 1. Luckily, the approach outlined in Section IV suggests a natural way to do so. In particular, it can be done by simply including a penalty term for the linear components of $y_k(\eta, Z)$ in addition to the penalty term needed to non-parametrically identify γ and δ. Note that this also corresponds to a case in the Bayesian framework outlined in Section III where we assume that: $[\alpha_k, \beta_k, \gamma_k]' \sim N(0, \sigma^2 I)$ for some finite σ, rather than only considering the limiting case where $\sigma^2 \to \infty$. See Mulhern et al. (2023) for more discussion of this point and an example of how it can work in practice.