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

Isaac M. Opper† Umut Özek‡

January 5, 2025

Abstract

We use a marginal treatment effect (MTE) representation of a fuzzy regression discontinuity setting to propose a novel estimation approach. The estimator can be thought of as extrapolating a 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 frequentist framework that it is consistent for the true MTE function 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 and the 2024 Economics of Education NBER Summer Institute 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

The past three decades have witnessed a surge in the use of regression discontinuity designs (RDDs) in applied economic research (Lee and Lemieux, 2010). Nowhere is this more apparent than in the economics of education, where the fact that many high-stakes educational decisions rely on student test scores with predetermined cutoffs has meant that RDDs could be used to evaluate a number of important policies.[1] Examples include admissions to selective or specialized public schools[2]; remedial or advanced course-taking in middle and high school[3]; high school graduation[4]; assignment to English learner, special education, or gifted education programs[5]; admission to elite colleges[6]; and many others. These decisions often incorporate exemptions to the test score rule, meaning that researchers use the discontinuity as an instrument to estimate the effects, i.e., use a fuzzy RDD for causal inference.

While fuzzy RDDs generally provide compelling evidence of a treatment effect, the resulting estimates are limited in the sense that the estimated effects only apply to students arbitrarily close to the discontinuity for whom being on the treatment side of the cutoff determines treatment status, i.e., the compliers at the cutoff (Bertanha and Imbens (2020)). Yet understanding treatment effects beyond compliers at the cutoff is almost always important from a public policy perspective. Doing so is required, for example, if one wants to determine whether explicit exemptions indeed identify students less likely to benefit from treatment or examine the effect of the treatment on students away from the cutoff. More generally, moving beyond the local average treatment effect (LATE) is important if one is interested in either assessing the overall effect of an existing policy or predicting how changing the policy would affect reading scores in the district.

In this study, we propose a new estimator, which we call Global RDD, for use in a fuzzy regression discontinuity setting and then use this approach to study how

[1] In fact, the study that introduced RDD examines the effects of merit awards on students' academic outcomes (Thistlethwaite and Campbell, 1960).

[2] e.g., Abdulkadiroğlu et al. (2014); Brunner et al. (2023); Clark (2010); Dobbie and Fryer (2011); Estrada and Gignoux (2017); Lucas and Mbiti (2014); Ozier (2018); Pop-Eleches and Urquiola (2013)

[3] e.g., Cortes and Goodman (2014); Figlio and Ozek (2024); McEachin et al. (2020); Özek (2021)

[4] e.g., Clark and Martorell (2014); Papay et al. (2022)

[5] e.g., Bui et al. (2014); Card and Giuliano (2016); Onda and Seyler (2020); Umansky (2016)

[6] e.g., Hastings et al. (2013); Zimmerman (2019)

global treatment effect estimates differ from LATE in an important education policy application: test-based grade retention. The Global RDD jointly models the two potential outcomes and selection into treatment using a marginal treatment effect (MTE) specification and then restricts the set of potential functions to ensure that our estimator converges to a unique MTE function. The result is an estimator that naturally builds on existing approaches used in other empirical contexts. For example, the estimator can be thought of as an extrapolation of the traditional fuzzy regression discontinuity estimate, where the 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 selection process does not vary away from the discontinuity. Alternatively, the estimator can 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).

We then 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. Using a Bayesian model in which the true conditional moments are themselves distributed according to some prior distribution, we also consider 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. Of note, we show that unless the researcher has strong priors in how the MTE deviate from being linear functions of the variables, the proposed estimator is nearly identical to the mean posterior if one is only interested in extrapolating to non-compliers and individuals whose scores are near (but not at) the discontinuity.

We then turn our attention to grade retention policies. Determining whether or not to retain a poorly performing student is one of the more important decisions that parents, teachers, and principals face regarding a child's education. On one hand, there is strong empirical evidence that retaining marginal students in early grades could increase their future test scores, especially if retention is coupled with additional instructional support. On the other hand, retaining a student is also quite costly, both in terms of the financial cost for the district and (potentially) the social cost for the individual student retained. This question is also particularly salient for education policy in the United States, since many states currently require school

districts to retain students whose test scores show that they are struggling to meet basic standards, although many of these states provide several exemptions.[7] While there is growing literature examining the effects of these policies using fuzzy RDDs (e.g., Greene and Winters (2007); Winters and Greene (2012); Özek (2015); Schwerdt et al. (2017); Figlio and Özek (2020); Hwang and Koedel (2022); Mumma and Winters (2023)), we know very little about their effects on students other than compliers right around the retention cutoffs.

Using student-level data from a large urban school district (LUSD) in Florida, we show that the benefits of retention are (1) larger for students with lower baseline reading achievement and (2) smaller for students exempt from retention. These findings imply that existing studies on early grade retention policies which rely on traditional fuzzy RD designs significantly underestimate the benefits of retention; for example, we show that the ATT estimates are roughly 20 percent larger than the LATE estimates on reading scores. This does not imply that more students should be retained, however, since we also find that 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), are negligible. We find that, as currently implemented, retaining students increases those students' sixth grade reading scores by 0.69σ, but further increasing the threshold by 50 points (roughly equivalent to moving the threshold from Level 2 to 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.

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

[7]For example, by 2020, about half of states required or encouraged school districts to retain students based on their third-grade reading scores. https://www.ecs.org/50-state-comparison-state-k-3-policies-2023/, accessed on 12/13/2024.

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.[8]

Letting T_i be a dummy variable denoting whether someone is in the treatment or control group, the observed outcome can be written as: $Y_i = \mu_i + \tau_i T_i$. We then follow the conventional specification of the MTE model and specify that the treatment is determined according to 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 researchers, we observe Y_i, T_i, and Z_i, but do not observe the latent variables μ_i, τ_i, and η_i nor do we observe the function $\nu(.)$.

Having specified the way individuals select into the treatment, we next assume that:

$$\mu_i = \mu^*(\eta_i, Z_i) + \epsilon_{\mu,i} \qquad \text{and} \qquad \tau_i = \tau^*(\eta_i, Z_i) + \epsilon_{\tau,i} \tag{1}$$

where the error terms are independently distributed according to some mean-zero distribution $(\epsilon_{\mu,i}, \epsilon_{\tau,i}) \sim F_i$. Non-random selection into treatment is therefore captured by the fact that both μ^* and τ^* can depend on the (unobserved) latent variable η_i, which also determines whether individual i selects into treatment. 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 to 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

[8] In Appendix B we discuss how to add additional covariates X_i to the analysis, but for expositional simplicity omit these for now.

related to what we observe. These moments are defined as follows:

$$y_0^*(\eta, Z) = \frac{1}{1-\eta}\int_\eta^1 \mu^*(\tilde{\eta}, Z)d\tilde{\eta} \qquad \text{and} \qquad y_1^*(\eta, Z) = \frac{1}{\eta}\int_0^\eta \mu^*(\tilde{\eta}, Z) + \tau^*(\tilde{\eta}, Z)d\tilde{\eta} \tag{2}$$

Note that these moments are redundant with the ones defined in Equation (1), i.e., if one knows the functions μ^* and τ^* it is clearly possible to calculate y_0^* and y_1^* and, similarly, if one knows y_0^* and y_1^* it is possible to compute μ^* and τ^*. For example, one can transform y_0 and y_1 to τ via the linear transformation:[9]

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

One of the main challenges in the regression discontinuity setting is that – even ignoring estimation error – we only observe the functions $y_0^*(\eta, Z)$ and $y_1^*(\eta, Z)$ at a select number of points. This means that making empirical statements about any estimand of interest other than the LATE requires, at a fundamental level, some extrapolation away from the observed conditional moments. To formally analyze this extrapolation, we will follow Opper (2024) and add a Bayesian hierarchical model on top of the traditional MTE framework. Specifically, we will assume that the true conditional moments are themselves distributed according to a Gaussian process. That is, we specify that:

$$y_0^*(\eta, Z) \sim \mathcal{GP}(\alpha_0 + \beta_0\nu + \gamma_0 Z, C_0) \tag{4}$$

$$y_1^*(\eta, Z) \sim \mathcal{GP}(\alpha_1 + \beta_1\nu + \gamma_1 Z, C_1) \tag{5}$$

$$[\alpha_0, \beta_0, \gamma_0, \alpha_1, \beta_1, \gamma_1]' \sim N(0, \sigma^2 I) \text{ with } \sigma^2 \to \infty \tag{6}$$

where $\mathcal{GP}(m, C)$ corresponds to a Gaussian process with mean m and covariance function C, and I corresponds to the identity matrix.

Since our main approach – as discussed below – will not be to estimate the fully Bayesian model, but to use the specification as a way to analyze the behavior of our estimator, we will not delve too deep into the intricacies of Gaussian processes; for the interested reader, Rasmussen and Williams (2006) provides an excellent intro-

[9] 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.

duction to Gaussian processes in general and Opper (2024) provides a more detailed discussions of Gaussian processes in the context of a marginal treatment effect model.

There are, however, some important notes about the model as specified in Equations (4) - (6). First, by specifying that the mean of the Gaussian processes are linear functions of η and Z – and using an uninformative prior limit for the linear terms – we specify that the baseline is a linear function of η and Z rather than function that equals zero at every point. In practice, this means that in the absence of more information the model extrapolates using a linear function of η and Z, rather than shrinking the predictions toward zero.

Second, the specification of the covariance functions C_k determine both the set of feasible functions as well as the relatively likelihood of functions within that set. For intuition, consider the case of a single parameters $\theta \in \mathbb{R}$ distributed according to some prior pdf $f(\theta)$. By specifying that $f(\theta) \sim N(0, \sigma^2)$ we would not place any bounds on the set of possible values for θ but specify that values close to zero are more likely than values far away from zero; in contrast, by specifying that $f(\theta) \sim U(\underline{\theta}, \overline{\theta})$, we would be assuming that $\theta \in (\underline{\theta}, \overline{\theta})$ while not specifying that any value within that set is more likely than any other. Similarly, based on our choice of C_1 we can restrict the set of functions y_1^* to be, for example, polynomial functions; in contrast, by specifying that C_1 is a squared-exponential kernel we can place very few restrictions on the set of potential functions y_1^*, but capture the intuition that smooth functions are more likely than functions which oscillate wildly.

Given this general framework, we now make the following three assumptions that we will hold throughout:

Assumption 1. $\eta_i | Z_i \sim U(0,1)$

Assumption 2. $\nu(Z) \in (0,1)$ *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 3. *The covariance functions* C_0 *and* C_1 *are both twice-differentiable functions of* (η, Z)*.*

Assumption 1 is relatively benign assumption and is common assumption in the MTE literature. The fact that η_i is distributed uniformly between zero and one is a

normalization as long as one is willing to assume that η_i is continuously distributed conditional on Z_i. This standard normalization is useful, as it implies that the cutoff value $\nu(Z_i)$ is equal to $Pr(T_i|Z_i)$, i.e. to the propensity score. This normalization also means that we can essentially think of η_i as being a ranking of how willing individual individual i is to opt-in to the treatment, relative to other individuals with the same value Z_i, with lower values indicating a higher willingness (or, equivalently, a lower cost).

Assumption 2 captures the fact that it is a fuzzy RD context, in that that for every value of the running variable there are both treated and untreated individuals (i.e., $\nu(Z) \in (0, 1)$) and that there is a single point Z^* at which the probability of treatment jumps discontinuously.[10] One advantage of the method is that it can be naturally extended to cases in which there are multiple discontinuities. For ease of exposition, however, we focus on the case with a single discontinuity and discuss in Appendix B how the method and results can be tweaked when there are multiple discontinuities.

Finally, by assuming that the covariance functions are twice-differentiable functions, Assumption 3 ensures that any realization of the Gaussian process (and hence the true potential outcome functions y_0^* and y_1^*) results in differentiable functions of η_i and Z_i. This can be thought of as a natural extension of the standard assumption required for RD designs that the potential outcome functions are continuous functions around the discontinuity, although it extends this assumption so that it applies globally (since we are interested in global effect estimates) and so that the functions are differentiable (so we can move from $y_k^*(\eta, Z)$ to $\tau(\eta, Z)$ using Equation (3)).

We conclude this section with one final assumption that we will relax in Section IV, which is listed below:

Assumption 4. *The researcher observes the conditional moments:* $\mathbb{E}[T_i|Z_i = Z, y_0^*, y_1^*]$, $\mathbb{E}[Y_i|T_i = 1, Z_i = Z, y_0^*, y_1^*]$, $\mathbb{E}[Y_i|T_i = 0, Z_i = Z, y_0^*, y_1^*]$ *at every point* $Z \neq Z^*$.

Assumption 4 specifies that the researcher observes the true conditional moments, rather than needing to estimate them. While apparently quite a strong assumption, this is meant to clarify the main ideas by allowing us to focus on questions of identification and extrapolation in the subsequent discussion of the estimator. We then

[10]The assumption also specifies that the probability increases as one moves across the threshold from left to right, but this is without loss of generality.

return to questions of estimation in Section IV; in that section, we show that if one replaces Assumption 4 with a common set of assumptions one 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 theoretical justification, as long as we take an asymptotic perspective.

Finally, note that the moments specified in Assumption 4 are conditional on the realization of the Gaussian process. Thus, we can view the expectations in the conventional frequentist context, where the true values are held as fixed and the expectations are taken over the error terms $\epsilon_{i,\mu}$, $\epsilon_{i,\tau}$, and η_i. In the rest of the paper we will generally leave this conditioning as implicit and simply write, for example, $\mathbb{E}[T_i|Z_i = Z]$. In cases where we do not condition on the realization of the Gaussian process, we will use the notation $\mathbb{E}_{\mathcal{GP}}$ to make explicit that the expectation is taken over the Gaussian process.

II.B Proposed Estimator

A natural next step is to estimate the Bayesian model outlined in Section II.A. This would involve more precisely specifying the covariance functions of the Gaussian process (and potentially some hyperpriors over the hyperparameters that govern these functions) and replacing Assumption 4 with assumptions about the sampling scheme, i.e., about F_i. It would then be possible to use the known characteristics of Gaussian processes to estimate posterior distributions of the MTE function and/or other estimands of interest.

While this approach has some advantages, it also has its downsides. In particular, the approach requires explicit specification of the priors (i.e., the covariance functions) and it is not entirely clear how the specification of the priors affects the resulting estimates. We therefore take an alternative approach in this paper. Instead of estimating the fully Bayesian model, we first propose an estimator and show that it is both feasible and well-defined. We then show that it translates the observed moments into the estimated MTE function in a natural way and one that can be thought of as naturally extending the existing alternatives. Finally, we explore both the conditions on the specified prior under which the resulting estimate converges to the true MTE function and the conditions under which it converges to the "best guess" of the true MTE function, i.e., the mean of the posterior, given our limited

set of observations.

With that outline, we 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:*

$$\tau^*_{GRDD} \equiv \hat{y}^*_1 - \hat{y}^*_0 + \eta \frac{\hat{y}^*_1}{\partial \eta} + (1-\eta)\frac{\partial \hat{y}^*_0}{\partial \eta} \tag{7}$$

where $\hat{y}^*_0$ *and* $\hat{y}^*_1$ *are defined such that for* $k \in \{0,1\}$*:*

1. $\hat{y}^*_k(\eta, Z) = \beta_k + \gamma_k(Z)$ *for some* $\beta_k \in \mathbb{R}$ *and continuous function* $\gamma_k(Z)$
2. $\hat{y}^*_k(\nu(Z), Z) = \mathbb{E}[Y_i | T_i = k, Z_i = Z]$ *for all* $Z \neq Z^*$

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. Second, τ^*_{GRDD} is well-defined, in that under the assumptions outlined in Section II.A there is guaranteed to be a single function τ^*_{GRDD} that meets that above definition. We state this as a proposition below:

Proposition 1. *Under the model outlined in Section II.A and Assumptions 1 - 4, the Global RDD as specified in Definition 1 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 and so even under the restriction 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. To see this, and to provide some intuition of how the Global RDD mechanically transforms the observed moments into the resulting estimates, we can consider a simplified example – illustrated in Figure 1 – 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 can start by looking at the relationship between the running variable – shown on the x-axis in Figure 1a – 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 and that this corresponds exactly to the function $\nu(Z)$. We indicate ten points on the function with dots, which we use in the other figures, and label six of them.

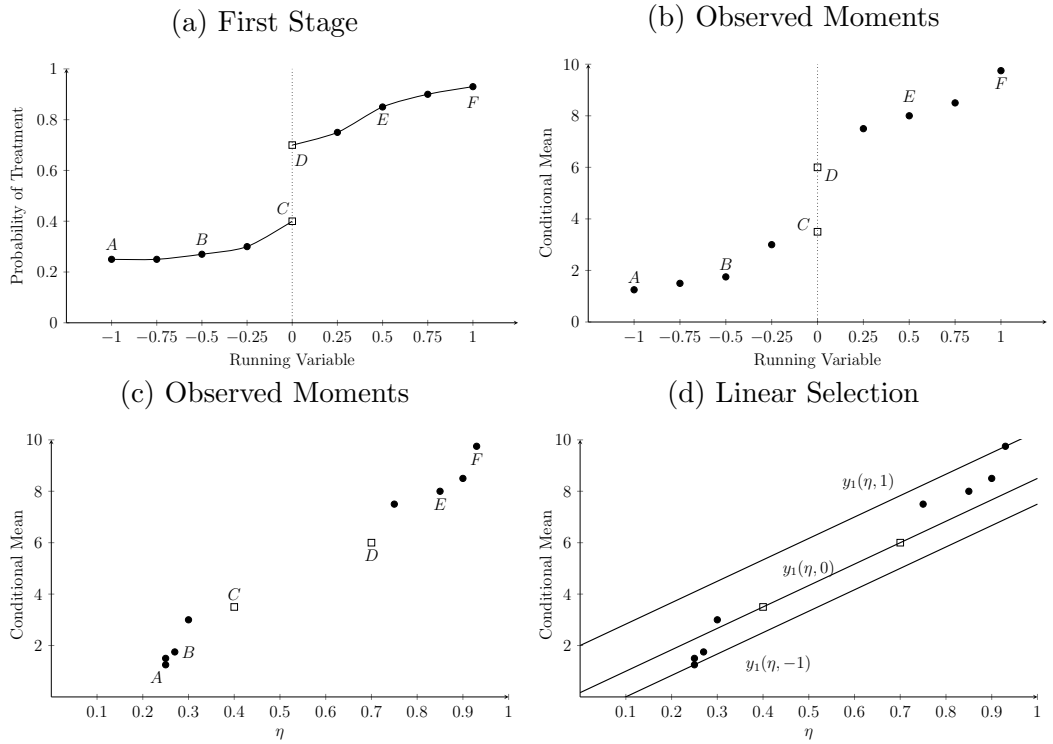
In the next two panels, we turn our attention to the observed conditional means, i.e., to $\mathbb{E}[Y_i|Z_i = Z, T_i = 1]$. We plot these observed moments for the ten points we highlighted in the previous panel in Figure 1b, labelling the same six points as in Figure 1a, with the running variable (i.e, 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 and so can think of these points we observe as being $y_1(\eta(Z), Z)$.

We can similarly plot the observed points with the value of η, rather than Z on the x-axis, i.e., as $y_1(\eta, \nu^{-1}(\eta))$. 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 $\hat{y}_1(\eta, 0)$. Note that this follows directly the discussion in Brinch et al. (2017). 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. Thus, the combination of: (a) the assumption that $y_1(\eta, 0)$ is a linear function of η and (b) $y_1(\eta, Z)$ is separable (i.e., the linear slope does not change away from the discontinuity) uniquely define the function $y_1(\eta, Z)$.

We conclude here by noting that we have not yet made any statements about how τ^*_{GRDD} is related the the true MTE function τ^* and only have shown that τ^*_{GRDD} is well-defined. There is, however, an intimate connection between the notion of an

Figure 1: Identification Intuition



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-axis. Panel (d) then shows how the Global RDD uses the information in these ten points to generate estimates of $\hat{y}_1(\eta, Z)$.

estimator being "well-defined" and a parameter being "point-identified," as both are fundamentally about there being a unique function that satisfies the restrictions. In fact, while we leave the discussion of this to Section IV, the fact that τ^*_{GRDD} is well-defined basically implies that the true MTE function τ^* is point identified if one is willing to assume that the conditional moments y_0 and y_1 are additively separable and linear in η.

II.C Alternative Definitions of the Estimator

In the previous section, we introduced the Global RDD and showed that it is both feasible and well-defined. In this section, we start by highlighting that even though the conditional moment functions in this estimator may be mis-specified, the proposed estimator gives the true local average treatment effect. Formally, we get thefollowing theorem:

Proposition 2. *Let τ^* denote the true MTE function, for any realization of the Gaussian process under any choice of C_0 and C_1. Then the estimated effect on the set of compliers at the Z^* is equal to the true effect on that set, i.e.:*

$$\frac{1}{p_h - p_l}\int_{p_l}^{p_h} \tau^*_{GRDD}(\eta, Z^*)d\eta = \frac{1}{p_h - p_l}\int_{p_l}^{p_h} \tau^*(\eta, Z^*)d\eta \tag{8}$$

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 simply use a traditional fuzzy RDD. In contrast to a traditional fuzzy RDD, however, the Global RDD also provides effect estimates away from the complier population at the cutoff. The question is therefore whether the particular extrapolation approach implied by the estimator defined in Section II is a good one.

Our formal answer to that question comes in the next section, where we discuss the restrictions on the covariance functions of the Gaussian processes and the treatment probabilities that either imply that the Global RDD converges to the true MTE function or (absent that) at least the best guess of true MTE function given the observed data. For another perspective, we now consider two natural alternatives: (1)

an observational study in which one compares the treatment average to the control average at every point $Z \in \mathbf{Z}$ and (2) a traditional fuzzy regression discontinuity design. Formally, we get that:

$$\tau^*_{obs}(Z) = \mathbb{E}[Y_i|T_i = 1, Z_i = Z] - \mathbb{E}[Y_i|T_i = 0, Z_i = Z] \tag{9}$$

$$\tau^*_{RDD} = \frac{1}{p_h - p_l}\left(\lim_{Z \downarrow Z^*} \mathbb{E}[Y_i|Z_i = Z] - \lim_{Z \uparrow Z^*} \mathbb{E}[Y_i|Z_i = Z]\right) \tag{10}$$

Here we show 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 to individuals away from the discontinuity. To do so, consider initially on the question of how extrapolate the LATE to other estimands of interest at the discontinuity. This is the same as extrapolating from the LATE to other estimands in an RCT or more general IV context. A natural approach is therefore to use the linear extrapolation approach as discussed in Brinch et al. (2017) and Kowalski (2023). Next, consider how to extrapolate away from the discontinuity; here, a natural approach to this extrapolation is to compare how the (potentially biased) observational study estimates vary, i.e., to compare $\tau^*_{obs}(Z)$ to $\tau^*_{obs}(Z^*_h)$. While formulated quite differently than Definition 1, as stated in the following Remark this is precisely how the Global RDD extrapolates away from the observed moments.

Remark 1. *Let τ^*_{GRDD} be the estimate generated by the Global Regression Discontinuity Design, as defined in Section II.B, and $\tau^*_{obs}(Z)$ and τ^*_{RDD} be the estimates generated from the traditional observational study and a traditional regression discontinuity design, as defined in Equation (9). Then if $\nu(Z) = p_h$, we get that:*

$$\tau^*_{GRDD}(\eta, Z) = \tau^*_{RDD} + extrap(\eta) + extrap(Z)$$

where

$$extrap(\eta) = 2 \cdot \left(\beta^*_1 - \beta^*_0\right) \cdot \left(\eta - \frac{p_h + p_l}{2}\right)$$
$$extrap(Z) = \tau^*_{obs}(Z) - \tau^*_{obs}(Z^*_h)$$

In a similar fashion, instead of starting with the traditional RDD we could start with a traditional observational study. A downside of such an observational study

is that we might be concerned that individuals endogenously choose (or are chosen) to enroll in the treatment. 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 use information at the discontinuity to do so, we first note that in the framework presented in Section II.A, endogenous selection stems from the fact that the conditional moments depend on the cost of enrollment, i.e., on η_i. 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 we could debias the observational study. Of course, this is not trivial; however, we can use the fact that – due to the discontinuity – we observe $y^*_1(\eta, Z^*)$ and $y^*_0(\eta, Z^*)$ at two different values of η to attempt such an adjustment.

Next, 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.,

$$\beta^*_0 \equiv \frac{y^*_0(p_h, Z^*) - y^*_0(p_l, Z^*)}{p_h - p_l} \tag{11}$$

$$\beta^*_1 \equiv \frac{y^*_1(p_h, Z^*) - y^*_1(p_l, Z^*)}{p_h - p_l} \tag{12}$$

Again, we can then use the linear extrapolation approach as discussed in Brinch et al. (2017) and Kowalski (2023) to adjust $\hat{y}_1(\nu(Z), Z)$ and $\hat{y}_0(\nu(Z), Z)$ based on β^*_0 and β^*_1, in hopes that it would improve the estimates that result from the observational study. Again, although specified differently this is an equivalent formulation of the Global RDD estimate. We specify this in the following Remark:

Remark 2. *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 (9). We then have:*

$$\tau^*_{GRDD}(\eta, Z) = \tau^*_{obs}(Z) - b \tag{13}$$

where b is a measure of the bias in the observational estimates. Specifically, defining

β_0^* and β_1^* as in Equation (11) and (12), we have:

$$b = \big(\nu(Z) - 2\eta\big) \cdot \big(\beta_1^* - \beta_0^*\big) - \beta_0^* \tag{14}$$

$$= \xi_h \cdot \big(\tau_{RDD}^* - \tau_{obs}^*(Z_h^*)\big) + \xi_l \cdot \big(\tau_{RDD}^* - \tau_{obs}^*(Z_l^*)\big) \tag{15}$$

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 a traditional fuzzy RDD, you could also think of it as an approach to debias a traditional observational study design. Second, we can write the bias term as a linear combination of the the traditional fuzzy RDD and observational study estimates at Z^*; see Equation (15). In other words, we can caluclate the expected bias in $\tau_{obs}^*(Z)$ by comparing the fuzzy RD estimate of the LATE to the $\tau_{obs}^*(Z)$ estimates at the discontinuity. The intuition for this formulation follows from the idea that if 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.[11] In contrast, if the fuzzy RD design diverges from them at the cutoff, this suggests that the observational estimates estimates are quite biased and therefore need significant bias adjustment.

Thus far, we have motivated our extrapolation approaches as being "natural" and/or being an application of earlier approaches. Using the Bayesian model specified in Section II.A, however, we can go beyond this motivation and show formally that the Global RDD indeed corresponds to adjusting the observational studies using "best guess" of the bias based on what is observed at discontinuity *regardless of the choice of prior*. Letting $\mathcal{D}(A)$ denote the observed conditional moments at points $Z \in A$, we state this in the following proposition:

Proposition 3. *Define to $b_{obs}^*(\eta, Z)$ be the bias in the observational study, i.e.,*

$$b_{obs}^*(\eta, Z) = \tau_{obs}^*(Z) - \tau^*(\eta, Z) \tag{16}$$

[11] 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.

where $\tau^*_{obs}(Z)$ is defined in Equation (9). Then:

$$\tau^*_{GRDD}(\eta, Z) = \tau^*_{obs}(Z) - \mathbb{E}_{\mathcal{GP}}\left[b^*_{obs}(\eta, Z)\middle|\mathcal{D}(\{Z^*, Z\})\right] \tag{17}$$

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

We build on this result in the next section, where we formally motivate the estimator.

III Comparing the GRDD to the True Treatment Effect Function

In the previous sections, we have introduced the model, proposed the estimator, showed that it was feasible and well-defined, and then discussed how it extrapolates away from the observed moments. In this section, we discuss conditions on the set of potential conditional moment functions and/or the prior likelihood of functions within this set that guarantees the proposed estimator either equals the true treatment effect function or is the "best guess", i.e., the mean of the posterior, of the true treatment effect function given the observed data.

III.A When does the GRDD Identify the True MTE Function?

As discussed at the end of Section II.B, the question of point identification is closely related to the question of whether the proposed estimator is well-defined. In fact, from the proof that the GRDD is well-defined, we can directly conclude that if the conditional moments are indeed additively separable and linear in η then the global RDD does equal the true MTE function. This result can be stated succinctly in the following proposition:

Proposition 4. *Suppose that $C_k = C_{k,Z}(Z, Z')$ for $k \in \{0, 1\}$. Then for any choice of $C_{k,Z}$, we have that $\tau^*_{GRDD}(\eta, Z) = \tau^*(\eta, Z)$.*

Absent any covariates, the assumption required for Proposition 4 is, in our opinion, relatively strong. With the addition of covariates, which we discuss how to do in Appendix B, the assumption may become much more tenable. Furthermore, 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.[12]

Proposition 5. *Letting τ_{AR} denote the Angrist and Rokkanen estimator, as defined in Angrist and Rokkanen (2015), we have the following two results:*

1. *Suppose C_k is such that $\tau^*_{obs}(Z) = \int_0^1 \tau^*(\eta, Z)d\eta$ for every realization of y^*. Then $\int_0^1 \tau^*_{GRDD}(\eta, Z) = \int_0^1 \tau^*(\eta, Z)d\eta$*
2. *Suppose C_k is such that $\tau^*_{AR}(Z) = \frac{1}{p_h - p_l}\int_{p_l}^{p_h} \tau^*(\eta, Z)d\eta$ for every realization of y^*. Then $\int_{p_l}^{p_h} \tau^*_{GRDD}(\eta, Z)d\eta = \int_{p_l}^{p_h} \tau^*(\eta, Z)d\eta$.*

While we leave the formal proof to Appendix A, we highlight here some nice similarity in the reasoning why the Global RDD relaxes the assumptions required for a traditional observational study and those required for the Angrist and Rokkanen method. In particular, the traditional observational study assumes that there is no endogenous 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, although it does put restrictions on how they do so.

It is worth highlighting that an important 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.* Similarly, the real advantage of an observational study is that it can be used even in the case where there is no treatment discontinuity. Thus, the proposition above stems in many ways from the fact that, in the specific context of a fuzzy RDD, neither approach is using the full set of observed

[12]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.

data, while the Global RDD does. Relaxing the necessary assumptions does have a cost; however, by relaxing the assumptions in the Global RDD the model is exactly identified and hence untestable. We discuss in Appendix B.C how the MTE formulation outlined in this paper also be used to specify intermediate models, which relax the assumptions in Angrist and Rokkanen and/or traditional observational studies but still allow for some of the assumptions to be testable.

III.B When is the GRDD the Best Guess?

Even though the Global RDD converges to the true MTE under weaker assumptions than the main alternative approaches, one may still not be comfortable with the assumptions required to identify the true MTE function. Motivated by the researchers' own experience, we therefore now consider in this 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. In particular, we now explore the conditions on the priors in which the Global RDD may still correspond to the researchers' "best guess" at the MTE function given the observed data, even if it is not necessarily the true MTE function.[13]

Much of the intuition for our analysis here stems from Proposition 3, which showed that the bias-adjustment implicit in the Global RDD corresponds to the expected bias based on what is observed at the discontinuity. An equivalent statement is that regardless of the choice of C_k the Global RDD corresponds to the mean posterior of the MTE function based on the moments observed at the discontinuity and one additional point. Again letting $\mathcal{D}(A)$ denote the observed conditional moments at points $Z \in A$, this proposition is stated below:

Proposition 6. *For any choice of C_0 and C_1 and any point $\tilde{Z} \neq Z^*$, we get that:*

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

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

This result seems quite promising; however, we observe data at all points $Z \in \mathbf{Z}$ and not just at two points and, unfortunately, Proposition 6 does not easily extend to

[13]Focusing on the mean posterior of the MTE function is important in part because the mean posterior is the optimal estimator in a Bayesian decision theory model with a squared-loss function.

show that $\tau^*_{GRDD}(\eta, Z) = \mathbb{E}_{\mathcal{GP}}\big[\tau^*(\eta, Z)|\mathcal{D}(\mathbf{Z})\big]$. The reason is that it is the covariance functions that determine whether one should 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 instead ascribes all such deviations to non-linearities in the relationship between Z and y_k, which may not align with the covariance functions.

We therefore consider conditions under which we can indeed interpret the Global RDD as the mean posterior conditional on all the observed data. Of course, one such case is when we should ascribe all deviations to non-linearities in the relationship between Z and y_k, which occurs when $C_k = C_{k,Z}(Z, Z')$. But the Global RDD also corresponds to the mean posterior if the treatment thresholds do not vary away from the discontinuity and one is willing to assume that the functions $y^*_k(\eta, Z)$ are separable, e.g., that the way selection into treatment occurs does not vary based on Z. 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:*

$$\nu(Z) = \begin{cases} p_l \text{ if } Z < Z^* \\ p_h \text{ if } Z > Z^* \end{cases} \tag{19}$$

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

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

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. We state this formally below for the conditional average treatment effects:

Proposition 8. *Let* $\tau^*_{GRDD}(Z) = \int_0^1 \tau^*_{GRDD}(\eta, Z)d\eta$ *and* $\tau^*(Z) = \int_0^1 \tau^*(\eta, Z)d\eta$*. Then for any* $\epsilon > 0$*, there exists a* $\delta > 0$ *such that for all* $Z \in (Z^* - \delta, Z^* + \delta)$*:*

$$\Big|\tau^*_{GRDD}(Z) - \mathbb{E}_{\mathcal{GP}}\big[\tau^*(Z)|\mathcal{D}(Z^* - \delta, Z^* + \delta)\big]\Big| < \epsilon \tag{21}$$

for every (η, Z).

IV Estimation, Inference, and Simulation Results

So far, we have assumed that the researchers observe the true conditional moment functions, i.e., $\mathbb{E}[Y_i|T_i = k, Z_i = Z]$ 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. Here, we first outline our estimation approach and additional assumptions, discuss convergence rates and the asymptotic distribution, and then highlight some important implementation details. Finally, we conclude with a simulation that shows the theoretical justifications of the Global RDD described above are backed up by impressive finite sample performance.

IV.A Estimation Approach, Sampling Assumptions, and the Limiting Distribution

There is a vast literature on the non-parametric estimation of conditional moment functions and our goal here is not to provide a detailed discussion of various non-parametric estimation approaches. Rather, we aim to outline an estimation approach and highlight how results from the existing literature can provide insights on the limiting distribution of the Global RDD.

To do so, we start by outlining a general estimation approach that uses linear smoothers to estimate the conditional moments and then uses the resulting estimates of these moments to estimate the MTE function. For the latter, we use the formulation of the Global RDD that specifies it is equivalent to a bias-adjusted observational study, as outlined in Proposition 2.

Our focus on linear smoothers, means that we estimate the conditional moment functions as:

$$\hat{\mathbb{E}}[Y_i|Z_i = Z, T_i = k] = \sum_{\forall i} \omega_k(Z, Z_i, T_i) \cdot Y_i \tag{22}$$

$$\hat{\mathbb{E}}[T_i|Z_i = Z] = \sum_{\forall i} \omega_\nu(Z, Z_i) \cdot T_i \tag{23}$$

for some set of weighting functions ω_0, ω_1, and ω_ν. Linear smoothers include most conventional non-parametric (and parametric) estimators, including kernel regres-

sions, smoothing splines, and series estimation. We assume that the weight functions depend on a tuning parameter, which measures the complexity of the model and which we denote as λ_0, λ_1, and λ_ν. For example, in series or basis estimation these λ's measure the number of terms included in the linear specification, while in kernel regressions they capture the inverse of the bandwidth. We allow the λ's (and hence the weight functions) to vary based on the sample size and allow them to depend on the full set of observed values $Z^n = (Z_1, ..., Z_n)$ and $T^n = (T_i, ..., T_n)$ as well; for notational convenience we omit this dependence when we write the weights.[14] It is useful to also define the following weighting functions, which are derived from the ones specified above:

$$\omega_{obs}(Z, Z_i, T_i) = \omega_1(Z, Z_i, T_i) - \omega_0(Z, Z_i, T_i) \tag{24}$$

$$\omega_1(\Delta Z^*, Z_i, T_i) = \lim_{Z \uparrow Z^*} \omega_1(Z, Z_i, T_i) - \lim_{Z \downarrow Z^*} \omega_1(Z, Z_i, T_i) \tag{25}$$

with $\omega_0(\Delta Z^*, Z_i, T_i)$ and $\omega_\nu(\Delta Z^*, Z_i)$ defined similarly.

Given these definitions, we define the Global RDD estimator for a general linear smoother as follows:

Definition 2. *Define* $\hat{\tau}_{GRDD}$ *as:*

$$\hat{\tau}_{GRDD}(\eta, Z) = \hat{\tau}_{obs}(Z) - \big(\hat{\nu}(Z) - 2\eta\big) \cdot \big(\hat{\beta}_1 - \hat{\beta}_0\big) - \hat{\beta}_0 \tag{26}$$

where:

$$\hat{\tau}_{obs}(Z) = \sum_{\forall i} \omega_{obs}(Z, Z_i, T_i) \cdot Y_i \tag{27}$$

$$\hat{\nu}(Z) = \sum_{\forall i} \omega_\nu(\Delta Z^*, Z_i) \cdot T_i \tag{28}$$

$$\hat{\beta}_1 = \frac{\sum_{\forall i} \omega_1(\Delta Z^*, Z_i, T_i) \cdot Y_i}{\sum_{\forall i} \omega_\nu(\Delta Z^*, Z_i) \cdot T_i} \tag{29}$$

$$\hat{\beta}_0 = \frac{\sum_{\forall i} \omega_0(\Delta Z^*, Z_i, T_i) \cdot Y_i}{\sum_{\forall i} \omega_\nu(\Delta Z^*, Z_i) \cdot T_i} \tag{30}$$

[14] A key assumption is that the weighting functions do not depend on the full set of outcomes $Y^n = (Y_1, ..., Y_n)$. In practice, the weighting functions often include tuning parameters which are chosen via cross-validation, which mean that they do depend on the Y^n. We ignore this complication here.

Having defined the estimator, we now turn to the sampling scheme. As specified in Section II.A, uncertainty comes from both the realization of the Gaussian process and from the error terms: $\epsilon_{\mu,i}$, $\epsilon_{\tau,i}$, and ν_i. However, we will focus here on conventional confidence intervals and condition on the realization of the Gaussian process.[15] By doing so, we are (roughly speaking) focusing here on how $\hat{\tau}_{GRDD}$ (as defined by Definition 2) compares to τ^*_{GRDD} (as defined by Definition 1), while the previous section was focused on how τ^*_{GRDD} compares to the true MTE function τ^*.

In Section II, we assume that $\epsilon_{\mu,i}$, $\epsilon_{\tau,i} \sim F_i$ are distributed independently across individuals as is $\eta_i \sim U(0,1)$. To show that the linear smoothers converge to their true values, we need a variety of additional regularity conditions, e.g., finite higher-order moments, sufficiently flexible nonparametric estimation, and enough range in the observed values Z^n to ensure a large number of observations near the value of Z we are interested in. Since these regularity conditions are well known, but vary based on the specific estimator, we will follow Kennedy (2023), and simply assume that estimators exhibit characteristics common to linear smoothers. These are specified below in Assumptions 5 and 6, in which we follow convention and write that $a \lesssim b$ if $a \leq c \cdot b$ for some constant c that does not depend on the sample size.

Assumption 5. *The weights on the linear smoothers are localized, in that:*

$$\omega_\nu(Z, Z_i) = 0 \text{ if } |Z_i - Z| \lesssim \cdot\lambda_\nu^{-1} \tag{31}$$

$$\omega_k(Z, Z_i, T_i) = 0 \text{ if } |Z_i - Z| \lesssim \cdot\lambda_k^{-1} \text{ for all } T_i \tag{32}$$

for $k \in \{0,1\}$ and for all $Z \neq Z^*$.

Assumption 6. *The linear smoothers exhibit the following bounds on their bias and variance:*

$$\Big|\mathbb{E}[\hat{\nu}(Z)] - \nu(Z)\Big| \lesssim \lambda_\nu^{-1} \qquad \mathbb{V}\big(\hat{\nu}(Z)\big) \lesssim \lambda_\nu \cdot n^{-1} \tag{33}$$

$$\Big|\mathbb{E}[\hat{y}_k(\nu(Z), Z)] - y_k^*(\nu(Z), Z)\Big| \lesssim \lambda_k^{-1} \qquad \mathbb{V}\big(\hat{y}_k(\nu(Z), Z)\big) \lesssim \lambda_k \cdot n^{-1} \tag{34}$$

for $k \in \{0,1\}$ and for all $Z \neq Z^*$.

These conditions are the same as specified in Kennedy (2023) and hold for a number of common estimators under standard regularity conditions. Consider, for

[15] We will also condition on the vector of observed Z_i's, i.e., on Z^n.

example, the Nadaraya-Watson estimator of $\hat{\nu}(Z)$ when using a uniform kernel with bandwidth $h_k = \lambda_k^{-1}$. Assumption 5 then clearly holds and it is straightforward to show that the bounds in Assumption 6 hold if we further assume that $\nu(Z)$ is continuously differentiable at all $Z \neq Z^*$. A more detailed discussion of the regularity conditions for which these hold for other linear smoothers can be found in Belloni et al. (2015) and Tsybakov (2009).

As is clear from the example of the Nadaraya-Watson estimator, the trick is to choose complex enough models to ensure there is minimal bias (i.e., λ_k is large), while not too complex so the variance is as small as possible (i.e., $\lambda_k \cdot n^{-1}$ is small). This tradeoff is highlighted in the proposition below, which both specifies the conditions needed for the resulting estimate to be asymptotically normal and computes the asymptotic variance.

Proposition 9. *Let $\hat{\tau}_{GRDD}$ be the estimate generated by the Global Regression Discontinuity Design, as defined in Definition 2, and τ^*_{GRDD} be the function defined in Definition 1. Then under Assumptions 5 and 6 and the choice of tuning parameters such that $\lambda_k \to \infty$ and $\lambda_k \cdot n^{-1} \to 0$ for $k \in \{0, 1, \nu\}$, we have that:*

$$\hat{\tau}_{GRDD}(Z, \eta) \xrightarrow{p} \tau^*_{GRDD}(Z, \eta) \tag{35}$$

for all η and $Z \neq Z^*$. Further, if we assume that $\omega_k(Z, Z_i, k') = 0$ if $k \neq k'$, $\lambda_k = \lambda$ for all k and that $\lambda^{-1}\sqrt{\lambda^{-1}n} \to 0$, we get that:

$$\sqrt{\lambda^{-1}n} \cdot \Big(\hat{\tau}_{GRDD}(Z, \eta) - \tau^*_{GRDD}(Z, \eta)\Big) \to N(0, V) \tag{36}$$

where the variance is equal to:

$$\begin{aligned} V = {} & \mathbb{V}\big(\hat{\tau}_{obs}(Z)\big) + \mathbb{V}\big(\hat{\nu}(Z)\big) \cdot \big(\beta_1^* - \beta_0^*\big)^2 \\ & + \mathbb{V}\big(\hat{\Delta}Y_1\big) \cdot \left(\frac{\nu(Z) - 2\eta}{\Delta p}\right)^2 + \\ & + \mathbb{V}\big(\hat{\Delta}Y_0\big) \cdot \left(\frac{\nu(Z) - 2\eta + 1}{\Delta p}\right)^2 \\ & + \mathbb{V}\big(\hat{\Delta}p\big) \cdot \left(\frac{\big(\nu(Z) - 2\eta\big) \cdot \big(\beta_1^* - \beta_0^*\big) - \beta_0^*}{\Delta p}\right)^2 \end{aligned} \tag{37}$$

and where:

$$\hat{\Delta}Y_1 = \sum_{\forall i} \omega_1(\Delta Z^*, Z_i, T_i) \cdot Y_i \tag{38}$$

$$\hat{\Delta}Y_0 = \sum_{\forall i} \omega_0(\Delta Z^*, Z_i, T_i) \cdot Y_i \tag{39}$$

$$\hat{\Delta}p = \sum_{\forall i} \omega_\nu(\Delta Z^*, Z_i) \cdot T_i \tag{40}$$

This proposition contains multiple results. First, as mentioned above it formalizes the conditions required for the estimate to be consistent and asymptotically normal. More interestingly, it also highlights that Assumption 5 implies that, asymptotically at least, the the estimate of $\hat{\tau}_{obs}(Z)$ is independent from the estimate of the bias terms $\hat{\beta}_1$ and $\hat{\beta}_0$; this is because the estimate of $\hat{\tau}_{obs}(z)$ relies on observations increasingly close to Z while the estimates of $\hat{\beta}_1$ and $\hat{\beta}_0$ rely on observations increasingly close to Z^*. Finally, in the expression of the the asymptotic variance it also shows that the estimates $\hat{\tau}_obs(Z)$ and $\hat{\nu}(Z)$ are also independent.

Proposition 9 is quite useful, but the parameter of interest is rarely a single point on the MTE function and instead an average over many values of the MTE function. For example, in the empirical example below we are interested in the effect of being retained on the test scores of those students who do not receive an exemption, i.e., we are interested in the average treatment effect on the treated (ATT). As we show in the proposition below, the asymptotic variance for those estimands is dominated by the bias terms: i.e., the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1 - \hat{\beta}_0$, rather than the estimates of $\hat{\tau}_{obs}$ and $\hat{\nu}(Z)$. This also highlights that the Global RDD is estimated at the same rate of convergence as a traditional RDD, with the differences being due to the different weights they put on $\hat{\Delta}Y_1$, $\hat{\Delta}Y_0$, and $\hat{\Delta}p$.

Proposition 10. *Let $\hat{\tau}_{GRDD}$ be the estimate generated by the Global Regression Discontinuity Design, as defined in Definition 2, and τ^*_{GRDD} be the function defined in Definition 1. Next, suppose that $\int_0^1 \omega(\eta, Z)d\eta = \omega_z(Z)$ for some continuous $\omega_z(Z) \in [0, \infty)$ with $\int_{\mathbf{Z}} \omega_z(Z) = 1$.*

Then under the assumptions in Proposition 9, we get that:

$$\sqrt{\lambda^{-1}n} \cdot \left(\int_{\boldsymbol{Z}} \int_0^1 \hat{\tau}_{GRDD}(\eta, Z)\omega(\eta, Z)d\eta dZ - \int_{\boldsymbol{Z}} \int_0^1 \tau^*_{GRDD}(\eta, Z)\omega(\eta, Z)d\eta dZ \right) \to N(0, V) \tag{41}$$

where the variance is equal to:

$$\begin{aligned} V = \mathbb{V}\big(\hat{\Delta}Y_1\big) \cdot \left(\frac{\xi}{\Delta p}\right)^2 + \mathbb{V}\big(\hat{\Delta}Y_0\big) \cdot \left(\frac{1+\xi}{\Delta p}\right)^2 \\ + \mathbb{V}\big(\hat{\Delta}p\big) \cdot \left(\frac{\xi \cdot \big(\beta_1^* - \beta_0^*\big) - \beta_0^*}{\Delta p}\right)^2 \end{aligned} \tag{42}$$

and where $\xi = \int_{\boldsymbol{Z}} \int_0^1 \big(\nu(Z) - 2\eta\big)\omega(\eta, Z) d\eta dZ$.

IV.B Implementation Details

We estimate the conditional moments required for the Global RDD using penalized smoothing splines. Formally, letting $Z_i^* = \mathbf{1}(Z_i \geq Z^*)$, this means we estimate $\hat{\nu}(Z)$ as $\hat{\nu}_0(Z) \cdot (1 - Z_i^*) + \hat{\nu}_1(Z) \cdot Z_i^*$, where:

$$\begin{aligned} \hat{\nu}_0, \hat{\nu}_1 = \underset{\nu_0, \nu_1}{\operatorname{argmin}} \Bigg\{ \sum_{\forall i} \Big(T_i - Z_i^* \cdot \nu_1(Z_i) - (1 - Z_i^*) \cdot \nu_0(Z_i)\Big)^2 \\ + \kappa_\nu \int \Big(\nu_0''(Z) + \nu_1''(Z)\Big)^2 dZ \Bigg\} \end{aligned} \tag{43}$$

Similarly, we estimate:

$$\begin{aligned} \hat{\alpha}_0, \hat{\alpha}_1, \hat{\gamma}, \hat{\Delta} = \underset{\alpha_0, \alpha_1, \gamma, \Delta}{\operatorname{argmin}} \Bigg\{ \sum_{\forall i} \Bigg(Y_i - \Big(Z_i^* \cdot \alpha_0(Z_i) + (1 - Z_i^*) \cdot \alpha_1(Z_i) \\ + T_i \cdot \gamma(Z_i) + T_i \cdot \Delta \cdot Z_i^*\Big)\Bigg)^2 \\ + \kappa_\alpha \int \Big(\alpha_0''(Z) + \alpha_1''(Z)\Big)^2 dZ + \kappa_\gamma \int \Big(\gamma''(Z)\Big)^2 dZ \Bigg\} \end{aligned} \tag{44}$$

where $\alpha(Z_i) = \alpha_0(Z_i) \cdot (1 - Z_i^*) + \alpha_1(Z_i) \cdot Z_i^*$, and then use these to construct estimates of $\hat{\tau}_{obs}$, $\hat{\beta}_0$, and $\hat{\beta}_1 - \hat{\beta}_0$, for use in Equation (71).[16]

There are a few important notes to make about this specification. Most importantly, rather than add a complexity penalty on the two conditional moments separately, we estimate the moments jointly using complexity penalties on $\alpha(Z)$ and

[16] In particular, $\hat{\tau}_{obs}(Z_i) = \hat{\gamma}(Z_i) + \hat{\Delta} \cdot Z_i^*$, $\hat{\beta}_0 = \frac{\alpha_1(Z^*) - \alpha_0(Z^*)}{\hat{\nu}_1(Z^*) - \hat{\nu}_0(Z^*)}$, and $\hat{\beta}_1 - \hat{\beta}_0 = \frac{\hat{\Delta}}{\hat{\nu}_1(Z^*) - \hat{\nu}_0(Z^*)}$

$\gamma(Z)$. In practice, this helps mitigate some issues that can arise when the distribution of the running variable is significantly different for the treated and control individuals, e.g., that the estimated difference between the two conditional moment estimates varies because of different amounts of smoothing that is used for the two conditional moment estimates. Note also that by specifying that $\hat{\tau}_{obs}(Z_i) = \hat{\gamma}(Z_i) + \hat{\Delta} \cdot Z_i^*$, we allow for $\hat{\tau}_{obs}(Z_i)$ to jump discontinuously at Z^* but require that its derivative, i.e., $\hat{\tau}'_{obs}(Z_i)$, is continuous around the discontinuity. In contrast, we do not make this assumption for $\alpha(Z)$ or $\nu(Z)$.

To implement this approach, we use the MGCV package in R. This automatically chooses the smoothing parameters κ_α, κ_γ, and κ_ν, specifies a low-rank approximation to the minimization problem, and can account for the selection of the κ terms in the estimated standard errors. See Wood (2017) for more information on the MGCV package and a primer on the theory behind this particular approach to linear smoothing. For inference, we do not use the asymptotic expressions of the variance specified in the section above, since this assumes the bandwidths are fixed and the locality restrictions may not hold in finite-samples. We do, however, use the delta method and the fact that the estimated propensity score function is uncorrelated with the estimates of the conditional moments.

Finally, we have developed an R package to implement this approach, which can be found at: https://github.com/isaacopper/GlobalRDD. The package also can account for clustering in the error terms when estimating the standard errors and (as discussed in Appendix B) allows for the addition of additional covariates and/or multiple discontinuities.

IV.C Monte Carlo Simulation

Before getting to the main empirical application, we conduct a simulation to compare the finite sample performance of the proposed approach (i.e., the Global RD design) with the two most natural alternatives – an observational study and a traditional RD design. The advantage of the simulation is that we can compare the estimated effect with the true effects, which allows us to generate estimates of the bias and mean-squared error (MSE) of the resulting estimates.

A main argument in the paper is that the proposed estimator is a valuable alternative to other approaches even if it is not necessarily a consistent estimator, e.g., if one

is not willing to assume that the true conditional moments are additively separable and linear in η. For our simulation, we therefore do not assume that the true conditional moments are linear in η or even additively separable in η and Z. Instead, we assume that the true conditional moments are generated by a non-separable Gaussian process where:

$$\mu^*(\eta, Z) = \alpha_\mu \eta + \beta_\mu Z + f_\mu(\eta) + g_\mu(Z) + h_{\mu,\eta}(\eta) \cdot h_{\mu,Z}(Z) \tag{45}$$

and $\alpha_\mu \sim N(0,1)$, $\beta_\mu \sim N(0,1)$, $f_\mu \sim \mathcal{GP}(0, C_{\mu,\eta})$, $g_Z \sim \mathcal{GP}(0, C_{Z,\eta})$, $h_{\mu,\eta} \sim \mathcal{GP}(0, C_{\mu,\eta})$, and $h_{\mu,Z} \sim \mathcal{GP}(0, C_{\mu,Z})$. For our covariance functions, we use a squared exponential with length scale 1 for $C_{\mu,\eta}$ and $C_{\tau,\eta}$ and length scale 2 for $C_{\mu,Z}$ and $C_{\tau,Z}$. We use an output variance of 1 for $C_{\mu,\eta}$ and $C_{\mu,Z}$ and of 0.5 for $C_{\tau,\eta}$ and $C_{\mu,Z}$ to capture the idea that the magnitude of treatment effects are (usually) a fraction of the overall variance in outcomes.

After generating the true conditional moments using the specification described in the previous paragraph, we then then simulate the data generating process by setting the sample size to be N and randomly determining each individuals' $Z_i \sim U(-1, 1)$ and $\eta_i \sim U(0, 1)$. We then determine their treatment status as:

$$T_i = \mathbf{1}\Big(\eta_i < \Phi\big(-0.75 + 0.75 \cdot Z + 0.75 \cdot \mathbf{1}(Z > -.5)\big)\Big)$$

where $\Phi(.)$ is the normal CDF and $\mathbf{1}$ represents the indicator function, and set their observed outcome to be:

$$Y_i = \mu(\eta_i, Z_i) + T_i \cdot \tau(\eta_i, Z_i) + \epsilon_i$$

where $\epsilon_i \sim N(0, 1)$.

Finally, given the generated data, we use the Global RDD approach outlined in this paper to estimate both the Average Treatment Effect (ATE) as well as the average effect on the compliers (LATE). We also estimate the LATE using a traditional RDD approach estimated via a local linear regression, both when using a fixed bandwidth of 0.5 as well as when estimating the bandwidth using the RDRobust command (Calonico et al. (2015a)). Finally, we also estimate the ATE using a propensity score weighting approach, both when estimating the propensity weights as well as when using the true propensity weights.

After each draw of the true conditional moments, we simulate the data and estimate the ATE/LATE 100 times. In doing so, we can separately estimate the squared bias – i.e., $\left(\tau - \mathbb{E}[\hat{\tau}]\right)^2$ – and the variance – i.e., $\mathbb{E}\left[\left(\hat{\tau} - \mathbb{E}[\hat{\tau}]\right)^2\right]$ for each set of true conditional moments, as well as the mean-squared error. We then do this for 100 randomly generated sets of true conditional moments and then report the average squared-bias, the average variance, and the average mean-square error for each of the estimators described above.

The results are shown in Table 1, which illustrate impressive performance by the Global RDD. In the top panel, we see that the Global RDD generates better estimates of the LATE than traditional RDD methods.[17] This does not stem from reduced bias – since the LATE is identified regardless of the true conditional moments, the squared-bias converges to (nearly) zero for both the traditional RDD and the Global RDD – but instead from lower variance. In other words, the estimation approach described above appears to do a better job (at least in our setting) of estimating the size of the discontinuity than using a locally linear approximation.

[17] In Table 1 we show the results of the local linear regression when using a fixed bandwidth, since it has lower mean-squared error than the RDRobust command. We show the results when using the RDRobust command to calculate the bandwidth in Appendix Table 6.

Table 1: Monte Carlo Results

(a) Estimates of LATE

	Squared Bias		Variance		Mean-Squared Error	
Sample Size	Local Lin.	Global RD	Local Lin.	Global RD	Local Lin.	Global RD
1,000	3.29	0.10	343.3	1.11	346.6	1.20
2,500	0.85	0.03	82.9	0.36	83.7	0.39
5,000	0.0022	0.01	0.26	0.17	0.26	0.19
10,000	0.0023	0.006	0.12	0.09	0.12	0.10

(b) Estimates of ATE

	Squared Bias		Variance		Mean-Squared Error	
Sample Size	PSWeight	Global RD	PSWeight	Global RD	PSWeight	Global RD
1000	0.85	0.27	0.006	5.88	0.86	6.16
2500	0.85	0.12	0.003	0.67	0.85	0.80
5000	0.85	0.10	0.001	0.25	0.85	0.35
10000	0.85	0.09	0.0006	0.13	0.85	0.13

Note: This table shows the results of the Monte Carlo simulation described in Section IV.C. In it, we generate 50 true conditional moments and for each of these, simulate the rest of the data generating process 100 times and each time estimate the treatment effects. We can then calculate the average squared-bias, variance, and mean-squared error. "Local Lin." is a traditional RDD estimate that uses a linear regression with triangle weights and a bandwidth of 0.5. "PSWeight" is a propensity score weighting approach using the true propensity scores. "Global RD" is the approach outlined in this paper. The simulation results with other alternatives are shown in Appendix Table 6.

In the bottom panel, we see that the Global RDD also generates better estimates of the ATE than traditional observational studies.[18] In contrast to the logic above, here the reduction stems not from increased precision and instead from reduced bias. Since we have not assumed that the true conditional moments are additively separable and linear in η, the squared-bias term does not disappear even as the sample size increases. However, making the "best-guess" adjustment for endogenous selection into the treatment does meaningfully reduce the bias relative to traditional observational studies, leading to better estimates of the ATE (i.e., lower mean-squared errors) when the sample size is sufficiently large.

Finally, note that the average variance is generally larger than the average squared-bias for the Global RDD, especially when the sample size is relatively small. This suggests that it may be worth tweaking the method to reduce the variance of estimates, even if doing so may add some (expected) bias. We discuss how to do so in Appendix B.D.

V The Effect of Grade Retention Policies

We now return to our empirical question about the effectiveness of grade retention policies, where fuzzy RDDs have become popular given the increasing use of student test score cutoffs to identify students to be retained. As we detail below, there is extensive literature examining the effects of grade retention on student outcomes using fuzzy RDDs; however, these estimated effects only apply to compliers (i.e., students not exempt from retention) right below retention cutoffs. We now use the Global RDD to determine whether the effects differ for exempt students and for lower-performing students identified for retention. We then use these results to better understand both the overall effect of the policy and if the policy could be designed to better improve outcomes.

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

[18] In Table 1 we show the results when we use the true propsensity score weights. We show the results when using the approach that estimates the propensity score in Appendix Table 6.

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. This policy requires students who score in the lowest achievement level on the 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 if they: (1) have been in the English learner program for less than two years; (2) have certain disabilities and have been already retained once until third grade; (3) have received intensive reading remediation for two years and have already been retained twice between kindergarten and third grade; (4) 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) 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)). Approximately half of the exemptions were due to the special education exemption and half due to the student showing proficiency either via an alternative test or via a teacher-developed portfolio, with the other two reasons contributing to a very small fraction of the exemptions.[19] While retention rates have gradually declined, in part due to improvements in reading achievement and in part due to an increase in exemption rates, they remain sizable and roughly 10 percent of the third graders were retained after the 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

[19]There are a large number of students in the English language learner program in Florida, but the majority had been in the program for more than two years by third grade and so do not qualify for the exemption.

(2020)). The overarching conclusion is that retained students outperform their same-age peers in the short term (one to three years), although these achievement gains partially fade out over time. Even with this fade-out, however, the studies generally find that 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 what follows, we present our 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.

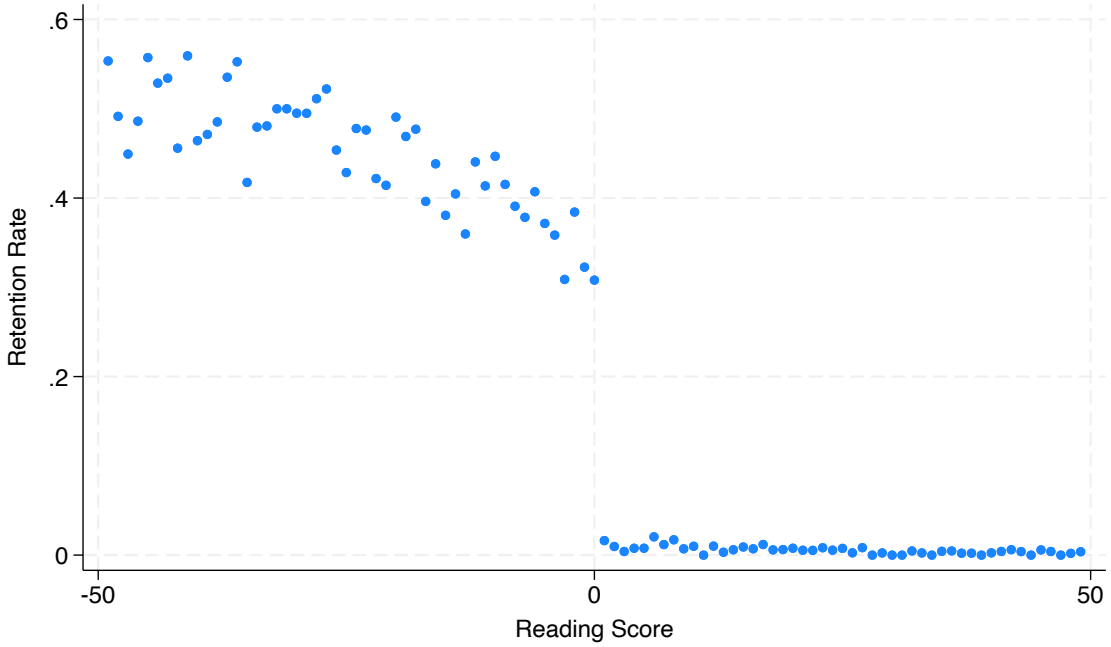
In our analysis, we use student-level administrative data from a large, urban school district (LUSD) in Florida. We use students who entered third grade for the first time between 2005-06 and 2010-11 school years and follow them until 8th grade. During this time period, the LUSD was approximately 30% White non-Hispanic, 30% Black, and 30% Hispanic, with approximately 60% of students on free/reduced price lunch. In our main analysis we restrict our attention to students without a disability, which corresponds to roughly 87% of the overall sample. This means we focus on the exemptions that are due to students' showing proficiency via alternative approaches and how the impact of retention varies away from the proficiency threshold. There are two reasons for this restriction: first, these exemptions are the most subject to debate; second, by focusing on a single type of exemption we lend more credibility to the MTE model, in which unobserved selection is captured by a single latent variable. Having said that, we find very similar results when we do not make this restriction; those results are shown in Appendix Table 5.

In the LUSD during our time period, roughly 13 percent of all students without a documented disability were flagged for retention and of those identified for retention, 44 percent were retained. Of those who were not flagged for retention, a small number of students who scored above the proficiency threshold were retained and so there is two-sided non-compliance in this setting.[20] Figure 2 shows the first stage, e.g., how the retention rate varies by third grade reading score. Our main outcomes of interest are standardized reading scores in grades 4 through 8.[21]

[20] Approximately 1% of students directly above the proficiency threshold were retained.

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

Figure 2: Probability of Being Retained



Note: The figure plots how the likelihood of being retained varies with the running variable. Each dot corresponds to the average retention rate of individuals with the same score on their third-grade reading test.

parison 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; however, we get similar results when doing a same-age comparison.

V.B Results

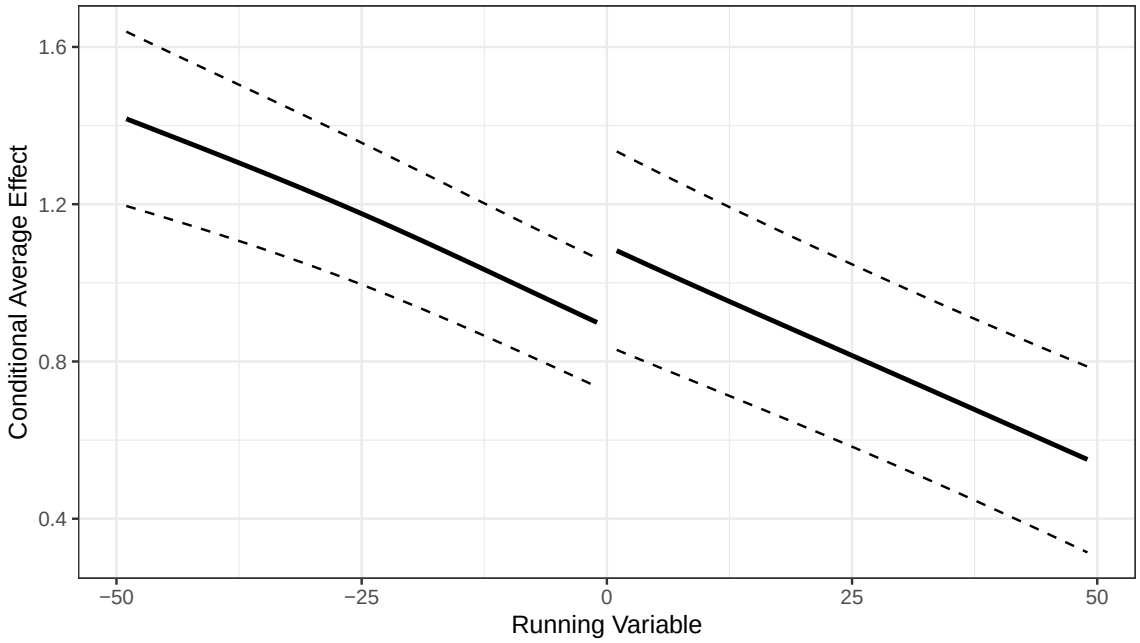
The overarching conclusion from the analysis is that the impact of retaining students is largest for those with lower third grade reading scores and for those who, conditional on their third-grade reading score, are most likely to be retained.

To illustrate these results, we start with Figure 3, which shows how the conditional average treatment effect on the treated individuals depends on their third grade reading scores. The solid line shows how the estimated conditional average treatment effect on the treated (CATT) – i.e., $\hat{\mathbb{E}}[\tau_i|Z_i = Z, T_i = 1]$ – varies based on the value of the running variable (third grade reading scale scores centered at the retention cutoff) and the dashed lines indicate the 95 percent point-wise confidence interval. The results in Figure 3 suggest that the positive effects of retention on fourth grade reading scores monotonically declines with students' baseline reading achievement. At the cutoff, we find that retention increases fourth grade reading scores by roughly 0.9σ. 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.

Figure 3 also highlights that at the discontinuity, the conditional average effect on the treated individuals jumps. This stems from the fact that the characteristics of the treated population discontinuously change at the threshold. We illustrate the effect of this more directly in Figure 4, which uses a contour map to show how $\hat{\tau}(\eta, Z)$ varies by both Z and η. In this exercise, η_i can be interpreted as a rank order of how likely an individual is to be retained: 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

Figure 3: Average Treatment Effect on the Treated

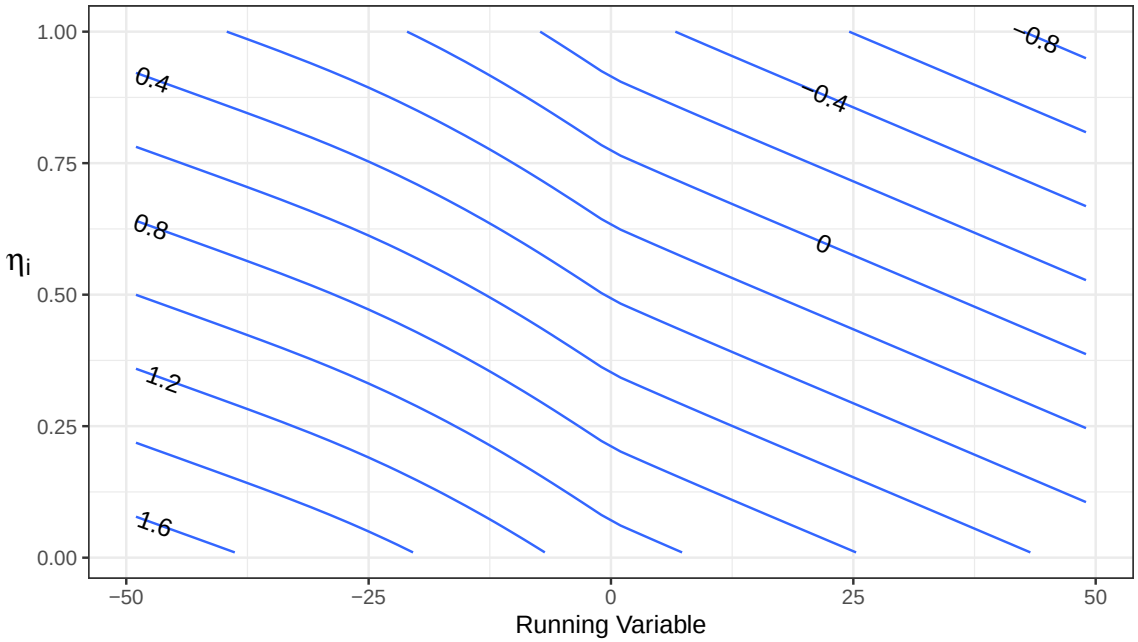


Note: The figure plots how the estimated the conditional average treatment effect on the treated varies with the running variable. Specifically, the solid line shows the estimated $\hat{\mathbb{E}}[\tau_i|Z_i = Z, T_i = 1]$ and the dashed lines indicated the 95% confidence interval, estimated via a Bayesian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

from retention. This finding suggests that the exemptions to the retention rule incorporated into Florida's policy indeed identify students who are least likely to benefit from retention.

As mentioned, the fact that the estimated effect is larger for students with lower reading test scores and for those with lower implied costs of retention (i.e., with lower η) suggests that the local average treatment effect, as identified by a traditional fuzzy RDD, underestimates the overall effect of the grade retention policy as it is currently implemented. To quantify this, we next use the estimates to calculate the estimated average effect of those retained under the realized retention decisions (i.e., the average effect of the treatment on the treated, or ATT), the estimated average effect on the complier population at the proficiency threshold (i.e., the LATE), and the estimated average effect if those with a reading score under 50 who were not retained under the current treatment assignment were in fact retained (i.e., the average treatment effect on the controls or ATC).

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



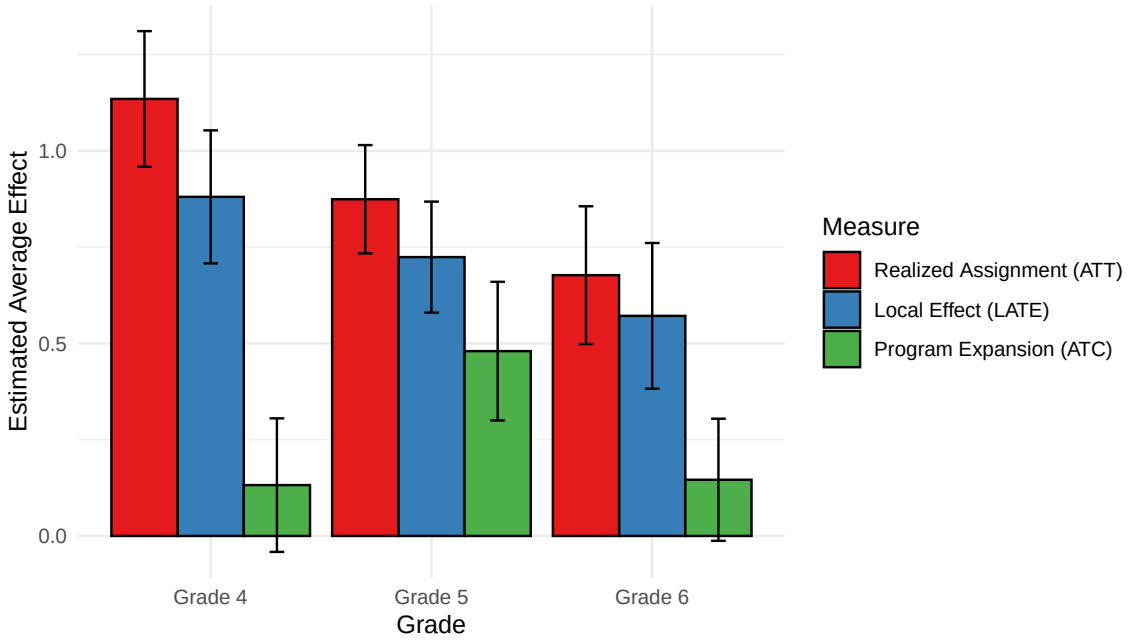
Note: The figure illustrates how $\hat{\tau}(\eta, Z) = \mathbb{E}[\tau_i|\eta_i = \eta, Z_i = Z]$ varies with both Z and η. Each line corresponds to a set of (η, Z) values with the same value of $\hat{\tau}(\eta, Z)$. Roughly speaking, η_i is a latent variable that serves as a measure of how likely 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.

We show these estimates for reading scores in grades 4, 5, and 6 in Figure 5. Focusing on fourth grade reading scores, the Global RDD estimates the LATE as being 0.88σ, which is consistent with the effect sizes found in the previous literature (Schwerdt et al. (2017), Figlio and Özek (2020)).[22] While this suggests that retention has a large positive effect on their outcomes, comparing the red bars to the blue bars in Figure 5 makes clear that the average effect on the compliers is consistently smaller than the estimated effect on all the students who are retained. For example, the estimated ATT for fourth grade reading scores is 1.14σ. While the error bars overlap, the two estimates are positively correlated with each other and we can reject

[22]This is also very similar to the LATE estimate that we get when using the rdrobust command on our data (Calonico et al., 2015b). This comparison is shown for all grade levels in Appendix Table 4.

(at the 1% level) the null hypothesis that the two estimates are equal each other.[23] Furthermore, the green bars shows that expanding the program by removing the exemptions and moving the proficiency threshold up 50 points would have much smaller effects.

Figure 5: Estimates of the ATT, LATE, and ATC



Note: The error bars correspond to the 95% confidence intervals. 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). 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).

In fact, we show in Appendix Table 2 that the realized assignment is nearly equivalent to the optimal assignment, defined as the policy that keeps the overall number of students retained constant but retains students with the highest predicted effect of retention. All of this suggests that Florida's policy is remarkably successful in identifying students most likely to benefit from retention.

[23]See Appendix Table 3 for estimates and confidence intervals of the differences between the estimates.

VI Conclusion

The trade-off between internal and external validity is a common issue in causal inference. Nowhere is this more clear than in a fuzzy regression discontinuity context, in which the local average treatment effect is identified under weak conditions but nearly all policy relevant treatment effects require an understanding of the effects on populations beyond compliers at the treatment discontinuity.

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. We show that it extends existing approaches in natural ways and can be motivated in both a frequentist 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 use this approach to examine the broader effects of early grade retention policies and show that understanding effects away from the discontinuity are important from a policy perspective. In particular, our findings suggest that the benefits of Florida's grade retention policy are larger than previously suggested, but that expanding the program would have a limited effect on the newly retained students.

We conclude by highlighting that the marginal treatment effect representation of the fuzzy RDD 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 discuss in Appendix B how the model can be extended to a number of important ways. In particular, the model can handle multiple discontinuities, incorporate additional covariances, provide alternative tests for the external validity of the traditional RDD estimates, and be used to improve the precision of the fuzzy RDD estimates.

References

Abdulkadiroğlu, Atila, Joshua Angrist, and Parag Pathak, "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," *Econometrica*, 2014, *82* (1), 137–196.

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.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato, "Some new asymptotic theory for least squares series: Pointwise and uniform results," *Journal of Econometrics*, 2015, *186* (2), 345–366.

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.

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

Brunner, Eric J., Shaun M. Dougherty, and Stephen L. Ross, "The Effects of Career and Technical Education: Evidence from the Connecticut Technical High School System," *The Review of Economics and Statistics*, 07 2023, *105* (4), 867–882.

Bui, Sa A., Steven G. Craig, and Scott A. Imberman, "Is Gifted Education a Bright Idea? Assessing the Impact of Gifted and Talented Programs on Students," *American Economic Journal: Economic Policy*, August 2014, *6* (3), 30–62.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik, "rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs," *R Journal*, 2015, *7* (1), 38–51.

_ , _ , and Rocío Titiunik, "rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs," *R Journal*, 2015, *7* (1), 38–51.

Card, David and Laura Giuliano, "Can Tracking Raise the Test Scores of High-Ability Minority Students?," *American Economic Review*, October 2016, *106* (10), 2783–2816.

Cattaneo, Matias D., 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.

Clark, Damon, "Selective Schools and Academic Achievement," *The B.E. Journal of Economic Analysis & Policy*, February 2010, *10* (1), 1–40.

_ and Paco Martorell, "The Signaling Value of a High School Diploma," *Journal of Political Economy*, 2014, *122* (2), 282–318.

Cortes, Kalena E. and Joshua S. Goodman, "Ability-Tracking, Instructional Time, and Better Pedagogy: The Effect of Double-Dose Algebra on Student Achievement," *American Economic Review*, May 2014, *104* (5), 400–405.

Dobbie, Will and Jr. Fryer Roland G., "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone," *American Economic Journal: Applied Economics*, July 2011, *3* (3), 158–87.

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.

Estrada, Ricardo and Jérémie Gignoux, "Benefits to elite schools and the expected returns to education: Evidence from Mexico City," *European Economic Review*, 2017, *95*, 168–194.

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.

_ **and Umut Ozek**, "The Unintended Consequences of Test-Based Remediation," *American Economic Journal: Applied Economics*, January 2024, *16* (1), 60–89.

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.

Hastings, Justine S, Christopher A Neilson, and Seth D Zimmerman, "Are Some Degrees Worth More than Others? Evidence from college admission cutoffs in Chile," Working Paper 19241, National Bureau of Economic Research July 2013.

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.

Kennedy, Edward H., "Towards optimal doubly robust estimation of heterogeneous causal effects," *Electronic Journal of Statistics*, 2023.

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.

Lucas, Adrienne M. and Isaac M. Mbiti, "Effects of School Quality on Student Achievement: Discontinuity Evidence from Kenya," *American Economic Journal: Applied Economics*, July 2014, *6* (3), 234–63.

McEachin, Andrew, Thurston Domina, and Andrew Penner, "Heterogeneous Effects of Early Algebra across California Middle Schools," *Journal of Policy Analysis and Management*, 2020, *39* (3), 772–800.

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.

Onda, Masayuki and Edward Seyler, "English learners reclassification and academic achievement: Evidence from Minnesota," *Economics of Education Review*, 2020, *79*, 102043.

Opper, Isaac M., "From LATE to ATE: A Bayesian Approach," *Journal of Econometrics*, 2024.

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

_ , "The effects of middle school remediation on postsecondary success: Regression discontinuity evidence from Florida," *Journal of Public Economics*, 2021, *203*, 104518.

Ozier, Owen, "The Impact of Secondary Schooling in Kenya," *Journal of Human Resources*, 2018, *53* (1), 157–188.

Papay, John P., Ann Mantil, and Richard J. Murnane, "On the Threshold: Impacts of Barely Passing High-School Exit Exams on Post-Secondary Enrollment and Completion," *Educational Evaluation and Policy Analysis*, 2022, *44* (4), 717–733.

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

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.

Tsybakov, Alexandre B., *Introduction to Nonparametric Estimation*, Springer Series in Statistics: Springer New York, NY, 2009.

Umansky, Ilana M., "Leveled and Exclusionary Tracking: English Learners' Access to Academic Content in Middle School," *American Educational Research Journal*, 2016, *53* (6), 1792–1833.

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.

Zimmerman, Seth D., "Elite Colleges and Upward Mobility to Top Jobs and Top Incomes," *American Economic Review*, January 2019, *109* (1), 1–47.

A Proofs

Proposition 1. *Under the model outlined in Section II.A and Assumptions 1 - 4, the Global RDD as specified in Definition 1 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. that is, we want to show that there is a single value $\beta_k \in \mathbb{R}$ and continuous function $\gamma_k(Z)$ that satisfies $\beta_k + \gamma_k(Z) = \mathbb{E}[Y_i|T_i = k, Z_i = z]$ 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 first highlight that although we do not observe the conditional moments at Z^*, any function $\hat{y}_0$ that is additively separately and linear in η and which satisfies the restriction that $\hat{y}_0(\nu(Z), Z) = \mathbb{E}[T_i = 0, Z_i = Z]$ for all $Z \neq Z^*$ also 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: $\hat{y}_0(\nu(Z^* + \epsilon), Z^* + \epsilon) = \mathbb{E}[Y_i|T_i = 0, Z_i = Z^* + \epsilon]$ and $\hat{y}_0(\nu(Z^* - \epsilon), Z^* - \epsilon) = \mathbb{E}[Y_i|T_i = 0, Z_i = Z^* - \epsilon]$ 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 $y_0^*(\eta, Z)$, there is a single choice of γ, defined as $\gamma^*(Z) = \mathbb{E}[Y_i|\nu(Z), Z, T_i = 0] - \beta_0^*\nu(Z)$, which satisfies $\hat{y}_k(\nu(Z), Z) = \mathbb{E}[Y_i|T_i = 0, Z_i = Z]$ for all $Z \neq Z^*$. Finally, from the assumption C_0 is differentiable guarantees that is that $\mu^*(\eta, Z)$ is continuous for all realizations of the Gaussian process and from the fact that $\nu(Z)$ and $\mu^*(\eta, Z)$ are both continuous functions, it follows that $\gamma^*(Z)$ is a continuous function.

□

Proposition 2. *Let τ^* denote the true MTE function, for any realization of the Gaussian process under any choice of C_0 and C_1. Then the estimated effect on the set of compliers at the Z^* is equal to the true effect on that set, i.e.:*

$$\frac{1}{p_h - p_l}\int_{p_l}^{p_h} \tau^*_{GRDD}(\eta, Z^*)d\eta = \frac{1}{p_h - p_l}\int_{p_l}^{p_h} \tau^*(\eta, Z^*)d\eta \tag{8}$$

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 both (p_h, Z^*) and (p_l, Z^*). The theorem thus follows. □

Remark 1. *Let τ^*_{GRDD} be the estimate generated by the Global Regression Discontinuity Design, as defined in Section II.B, and $\tau^*_{obs}(Z)$ and τ^*_{RDD} be the estimates generated from the traditional observational study and a traditional regression discontinuity design, as defined in Equation (9). Then if $\nu(Z) = p_h$, we get that:*

$$\tau^*_{GRDD}(\eta, Z) = \tau^*_{RDD} + extrap(\eta) + extrap(Z)$$

where

$$extrap(\eta) = 2 \cdot \left(\beta^*_1 - \beta^*_0\right) \cdot \left(\eta - \frac{p_h + p_l}{2}\right)$$
$$extrap(Z) = \tau^*_{obs}(Z) - \tau^*_{obs}(Z^*_h)$$

Proof. To start, we will note that using the restriction that $\hat{y}^*_k(\eta, Z) = \beta^*_k\eta + \gamma^*_k(Z)$ we can re-write Equation (7) to get that:

$$\begin{aligned}\tau^*_{GRDD}(\eta, Z) &= \left(\beta^*_1\eta + \gamma^*_1(Z)\right) - \left(\beta^*_0\eta + \gamma^*_0(Z)\right) + \eta\beta^*_1 + (1-\eta)\beta^*_0 \\ &= \gamma^*_1(Z) - \gamma^*_0(Z) + 2\eta \cdot (\beta^*_1 - \beta^*_0) - \beta^*_0\end{aligned} \tag{46}$$

Next, we can write both the traditional RDD and the observational study in terms of $y^*_k(\eta, Z)$ as follows:

$$\tau^*_{RDD} = \frac{\left(p_h \cdot y^*_1(p_h, Z^*) + (1-p_h) \cdot y^*_0(p_h, Z^*)\right) - \left(p_l \cdot y^*_1(p_l, Z^*) + (1-p_l) \cdot y^*_0(p_l, Z^*)\right)}{p_h - p_l} \tag{47}$$

$$\tau^*_{obs}(Z) = y^*_1(\nu(Z), Z) - y^*_0(\nu(Z), Z) \tag{48}$$

Finally, we get from the proof of Proposition 1 that β^*_0 and β^*_1 are defined such

that:

$$\beta_0^* = \frac{y_0^*(p_h, Z^*) - y_0^*(p_l, Z^*)}{p_h - p_l} \qquad \text{and} \qquad \beta_1^* = \frac{y_1^*(p_h, Z^*) - y_1^*(p_l, Z^*)}{p_h - p_l} \tag{49}$$

Combining these expressions, we thus get the following relationships between the different estimators:[24]

$$\tau_{GRDD}^*(\eta, Z) = \tau_{obs}^*(Z) - \big(\nu(Z) - 2\eta\big) \cdot (\beta_1^* - \beta_0^*) + \beta_0^* \tag{50}$$

$$\tau_{RDD}^* = \tau_{obs}^*(Z_h^*) + p_l \cdot (\beta_1^* - \beta_0^*) + \beta_0^* \tag{51}$$

where $\tau_{obs}^*(Z_h^*) \equiv y_1^*(p_h, Z^*) - y_1^*(p_h, Z^*)$.

If we now consider the case where $\nu(Z) = p_h$, we get that:

$$\begin{aligned}
\tau_{GRDD}^*(\eta, Z) &= \tau_{obs}^*(Z) - \big(p_h - 2\eta\big) \cdot (\beta_1^* - \beta_0^*) + \beta_0^* \\
&= \tau_{RDD}^* - \big(\tau_{obs}^*(Z_h^*) + p_l \cdot (\beta_1^* - \beta_0^*) + \beta_0^*\big) \\
&\qquad + \tau_{obs}^*(Z) - \big(p_h - 2\eta\big) \cdot (\beta_1^* - \beta_0^*) + \beta_0^* \\
&= \tau_{RDD}^* + \big(\tau_{obs}^*(Z) - \tau_{obs}^*(Z_h^*)\big) + (\beta_1^* - \beta_0^*) \cdot (2\eta - p_h - p_l)
\end{aligned}$$

Re-writing $2\eta - p_h - p_l = 2 \cdot (\eta - \frac{p_h + p_l}{2})$ and we get the result. □

Remark 2. *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 (9). We then have:*

$$\tau_{GRDD}^*(\eta, Z) = \tau_{obs}^*(Z) - b \tag{13}$$

where b is a measure of the bias in the observational estimates. Specifically, defining β_0^* and β_1^* as in Equation (11) and (12), we have:

$$b = \big(\nu(Z) - 2\eta\big) \cdot \big(\beta_1^* - \beta_0^*\big) - \beta_0^* \tag{14}$$

$$= \xi_h \cdot \big(\tau_{RDD}^* - \tau_{obs}^*(Z_h^*)\big) + \xi_l \cdot \big(\tau_{RDD}^* - \tau_{obs}^*(Z_l^*)\big) \tag{15}$$

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 $\tau_{obs}^*(Z_l^*) = \lim_{Z \uparrow Z^*} \tau_{obs}^*(Z)$.

[24] For the derivation of the expression for τ_{RDD}^* see the proof below.

Proof. The fact that we can write $\tau^*_{GRDD}(\eta, Z)$ as $\tau^*_{obs}(Z) - b$ for a bias term $b = \big(\nu(Z) - 2\eta\big) \cdot \big(\beta^*_1 - \beta^*_0\big) - \beta^*_0$ follows directly from the proof above. We will therefore focus here on showing that we can re-write the bias term as $\xi_h \cdot \big(\tau^*_{RDD} - \tau^*_{obs}(Z^*_h)\big) + \xi_l \cdot \big(\tau^*_{RDD} - \tau^*_{obs}(Z^*_l)\big)$ 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 $\tau^*_{obs}(Z^*_l) = \lim_{Z \uparrow Z^*} \tau^*_{obs}(Z)$.

As before, we start by re-writing Equation (47) to get that:

$$\begin{aligned}
\tau^*_{RDD} &= \frac{\big(p_h \cdot y^*_1(p_h, Z^*) + (1 - p_h) \cdot y^*_0(p_h, Z^*)\big) - \big(p_l \cdot y^*_1(p_l, Z^*) + (1 - p_l) \cdot y^*_0(p_l, Z^*)\big)}{p_h - p_l} \\
&= p_h \cdot \frac{y^*_1(p_h, Z^*) - y^*_0(p_h, Z^*)}{p_h - p_l} + \frac{y^*_0(p_h, Z^*) - y^*_0(p_l, Z^*)}{p_h - p_l} - p_l \cdot \frac{y^*_1(p_l, Z^*) - y^*_0(p_l, Z^*)}{p_h - p_l} \\
&= \frac{p_h}{p_h - p_l} \tau^*_{obs}(Z^*_h) + \beta^*_0 - p_l \cdot \frac{y^*_1(p_l, Z^*) - y^*_0(p_l, Z^*)}{p_h - p_l} \\
&= \tau^*_{obs}(Z^*_h) + \beta^*_0 - p_l \cdot \frac{y^*_1(p_l, Z^*) - y^*_0(p_l, Z^*)}{p_h - p_l} + p_l \cdot \frac{y^*_1(p_h, Z^*) - y^*_0(p_h, Z^*)}{p_h - p_l} \\
&= \tau^*_{obs}(Z^*_h) + \beta^*_0 + p_l \cdot \frac{y^*_1(p_h, Z^*) - y^*_1(p_l, Z^*)}{p_h - p_l} + p_l \cdot \frac{y^*_0(p_h, Z^*) - y^*_0(p_l, Z^*)}{p_h - p_l} \\
&= \tau^*_{obs}(Z^*_h) + \beta^*_0 + p_l \cdot \big(\beta^*_1 - \beta^*_0\big)
\end{aligned}$$

We can similarly show that we could also write τ^*_{RDD} as $\tau^*_{obs}(Z^*_l) + \beta^*_0 + p_h \cdot \big(\beta^*_1 - \beta^*_0\big)$. From this, we can set up the series of linear equations:

$$\begin{bmatrix} 1 - p_h & p_h \\ 1 - p_l & p_l \end{bmatrix} \begin{bmatrix} \beta^*_0 \\ \beta^*_1 \end{bmatrix} = \begin{bmatrix} \tau^*_{RDD} - \tau^*_{obs}(Z^*_l) \\ \tau^*_{RDD} - \tau^*_{obs}(Z^*_h) \end{bmatrix} \tag{52}$$

which makes it clear that we can write β^*_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 β^*_1. Plugging that into the expression that $b = \big(\nu(Z) - 2\eta\big) \cdot \big(\beta^*_1 - \beta^*_0\big) - \beta^*_0$ we get that we can also write:

$$\tau^*_{GRDD}(\eta, Z) = \tau^*_{obs}(Z) + \xi_h \cdot \big(\tau^*_{RDD} - \tau^*_{obs}(Z^*_h)\big) + \xi_l \cdot \big(\tau^*_{RDD} - \tau^*_{obs}(Z^*_l)\big) \tag{53}$$

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

□

Proposition 3. *Define to $b^*_{obs}(\eta, Z)$ be the bias in the observational study, i.e.,*

$$b^*_{obs}(\eta, Z) = \tau^*_{obs}(Z) - \tau^*(\eta, Z) \tag{16}$$

where $\tau^*_{obs}(Z)$ is defined in Equation (9). Then:

$$\tau^*_{GRDD}(\eta, Z) = \tau^*_{obs}(Z) - \mathbb{E}_{\mathcal{GP}}\big[b^*_{obs}(\eta, Z)\big|\mathcal{D}(\{Z^*, Z\})\big] \tag{17}$$

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

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:

$$\mathbb{E}\Big[y_k(\eta, Z)\big|\mathcal{D}\big(\{\tilde{Z}, Z^*\}\big)\Big] = h(\eta, Z)'\hat{\theta} + C_k\big((\eta, Z), \mathcal{D}\big)'C_k\big(\mathcal{D}\big)^{-1}r \tag{54}$$

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 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:

$$\mathbb{E}\Big[y_k(\eta, Z)\big|\mathcal{D}\big(\{\tilde{Z}, Z^*\}\big)\Big] = (\eta - \nu(Z))\beta^*_k + y_k\big(\nu(Z), Z\big) \tag{55}$$

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 the transformation from y to τ is linear, we therefore get that $\mathbb{E}[\tau^*(\eta, Z)|\mathcal{D}(\{\tilde{Z}, Z^*\})] = \tau^*_{GRDD}(\eta, Z)$ for all η and $Z \in \{\tilde{Z}, Z^*\}$. □

Proposition 4. *Suppose that $C_k = C_{k,Z}(Z, Z')$ for $k \in \{0, 1\}$. Then for any choice of $C_{k,Z}$, we have that $\tau^*_{GRDD}(\eta, Z) = \tau^*(\eta, Z)$.*

Proof. By restricting the covariance function to be $C_k = C_{k,Z}(Z, Z')$ for $k \in \{0, 1\}$, we can infer that the realization of the Gaussian process is additively separable and linear in η. The result then follows from the fact that, as shown in Proposition 1, the $\hat{y}$ resulting from the Global Regression Discontinuity Design are the only y that is both additively separable and linear in η, while also matching y^* at all of the observed moments. □

Proposition 5. *Letting τ_{AR} denote the Angrist and Rokkanen estimator, as defined in Angrist and Rokkanen (2015), we have the following two results:*

1. *Suppose C_k is such that $\tau^*_{obs}(Z) = \int_0^1 \tau^*(\eta, Z)d\eta$ for every realization of y^*. Then $\int_0^1 \tau^*_{GRDD}(\eta, Z) = \int_0^1 \tau^*(\eta, Z)d\eta$*
2. *Suppose C_k is such that $\tau^*_{AR}(Z) = \frac{1}{p_h - p_l}\int_{p_l}^{p_h} \tau^*(\eta, Z)d\eta$ for every realization of y^*. Then $\int_{p_l}^{p_h} \tau^*_{GRDD}(\eta, Z)d\eta = \int_{p_l}^{p_h} \tau^*(\eta, Z)d\eta$.*

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.[25] 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 $p_h(X) = \lim_{Z \downarrow Z^*} \nu(Z, X)$.[26] 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:

$$\tau^*_{GRDD}(\eta, Z, X) = \tau^*_{obs}(Z, X) - \beta^*_{0,X} \cdot \nu(Z, X) - \beta^*_{1,X} \cdot \big(1 - \nu(Z, X)\big) \tag{56}$$

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 $\beta^*_{1,X} = \frac{y_1(p_h(X), Z^*, X) - y_1(p_l(X), Z^*, X)}{p_h(X) - p_l(X)}$.

[25] 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 Appendix B.A.

[26] We extend Assumption 2 to be that $p_h(X) \neq p_l(X)$ for all 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:

$$y_k(\eta, Z, X) = y_k(\eta, X) \tag{57}$$

$$\nu(Z,X) = p_h(X) \cdot \mathbf{1}(Z_i > Z^*) + p_l(X) \cdot \mathbf{1}(Z_i < Z^*) \tag{58}$$

and the *Conditional first stage* being that $p_h(X) \neq p_l(X)$ for all X.[27] Under this assumption, it follows that:

$$\frac{1}{p_h(X) - p_l(X)} \int_{p_l(X)}^{p_h(X)} \tau^*_{GRDD}(\nu, X) d\nu = \frac{1}{p_h(X) - p_l(X)} \int_{p_l(X)}^{p_h(X)} \tau^*(\nu, X) d\nu \tag{59}$$

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

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

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

Proof. This follows directly from Proposition 3. □

[27] As shown in Vytlacil (2002), the monotonicity assumption is implicit in the generalized Roy model introduced in Section II.A.

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:*

$$\nu(Z) = \begin{cases} p_l \text{ if } Z < Z^* \\ p_h \text{ if } Z > Z^* \end{cases} \tag{19}$$

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

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

for every (η, Z)*.*

Proof. We start by re-casting mean posterior of the Gaussian process as:

$$\mathbb{E}\big[y_k(\eta, Z)\big|\mathcal{D}(\mathbf{Z})\big] = h(\eta, Z)'\hat{\theta} + \hat{\tilde{y}}(\eta, Z) \tag{60}$$

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

$$\arg\min_{\theta, y} ||y|| \text{ s.t. } h(\nu(Z), Z)'\theta + y(\nu(z), Z) = y_k(\nu(Z), Z)\ \forall Z \tag{61}$$

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 61 as:

$$\arg\min_{\beta_k, f} ||f|| \text{ s.t. } \beta_k\nu(Z) + f(\nu(Z)) = y_k(\nu(Z), Z) - \big(\alpha_k + \gamma_k Z + g(Z)\big)\ \forall Z \tag{62}$$

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:

$$\mathbb{E}\big[y_k(\eta, Z)\big|\mathcal{D}(\mathbf{Z})\big] = (\eta - \nu(Z))\beta^*_k + y_k\big(\nu(Z), Z\big) \tag{63}$$

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* $\tau^*_{GRDD}(Z) = \int_0^1 \tau^*_{GRDD}(\eta, Z)d\eta$ *and* $\tau^*(Z) = \int_0^1 \tau^*(\eta, Z)d\eta$. *Then for any* $\epsilon > 0$, *there exists a* $\delta > 0$ *such that for all* $Z \in (Z^* - \delta, Z^* + \delta)$:

$$\left|\tau^*_{GRDD}(Z) - \mathbb{E}_{\mathcal{GP}}\big[\tau^*(Z)|\mathcal{D}(Z^* - \delta, Z^* + \delta)\big]\right| < \epsilon \tag{21}$$

for every (η, Z).

Proof. We will show that for any $\epsilon > 0$ there exists a $\delta > 0$ such that:

$$|\hat{y}^*_k(\eta, Z) - \mathbb{E}_{GP}[y^*_k(\eta, Z)|\mathcal{D}(Z^* - \delta, Z^* + \delta)]| < \epsilon \tag{64}$$

for $k \in \{0, 1\}$ and all (η, Z). Clearly, the conclusion follows from this. To do so, we will use that fact that we can approximate $\mathbb{E}_{GP}[y^*_k(\eta, Z)|\mathcal{D}(Z^* - \delta, Z^* + \delta)]$ arbitrarily well using the formula:

$$\mathbb{E}_{GP}[y^*_k(\eta, Z)|\mathcal{D}(Z^* - \delta, Z^* + \delta)] \approx C_{y^*_k(\eta,Z),Y^{obs}}\big(C_{Y^{obs}} + \Sigma\big)^{-1}Y^{obs} \tag{65}$$

where Y^{obs} is an $M \times 1$ vector defined as $[y^*_k(\nu(Z^* - \delta), Z^* - \delta), y^*_k(\nu(Z^* - \delta + \Delta), Z^* - \delta + 2 * \Delta), Z^* - \delta + 2 * \Delta), ..., y^*_k(\nu(Z^* + \delta), Z^* + \delta)]$ for a small enough Δ and where $C_{y^*_k(\eta,Z),Y^{obs}}$ is the covariance function.

By choosing δ, we can make the weights arbitrarily close to an additively separable covariance and $\nu(Z)$ arbitrarily close to a step-function. Since the function $C_{y^*_k(\eta,Z),Y^{obs}}\big(C_{Y^{obs}}+\Sigma\big)^{-1}Y^{obs}$ is continuous in the weights, we can therefore – by choosing δ – make the resulting values of $\mathbb{E}_{GP}[y^*_k(\eta, Z)|\mathcal{D}(Z^* - \delta, Z^* + \delta)]$ arbitrarily close to what they would be if using an additively separable covariance function and if $\nu(Z)$ was a step-function. From the last proof, the value of $\mathbb{E}_{GP}[y^*_k(\eta, Z)|\mathcal{D}(Z^* - \delta, Z^* + \delta)]$ in that case is precisely the Global RDD estimates of $\hat{y}^*_k(\eta, Z)$. Thus, by choosing δ, we can make $\hat{y}^*_k(\eta, Z)$ arbitrarily close to $\mathbb{E}_{GP}[y^*_k(\eta, Z)|\mathcal{D}(Z^* - \delta, Z^* + \delta)]$. □

Proposition 9. *Let* $\hat{\tau}_{GRDD}$ *be the estimate generated by the Global Regression Discontinuity Design, as defined in Definition 2, and* τ^*_{GRDD} *be the function defined in Definition 1. Then under Assumptions 5 and 6 and the choice of tuning parameters such that* $\lambda_k \to \infty$ *and* $\lambda_k \cdot n^{-1} \to 0$ *for* $k \in \{0, 1, \nu\}$, *we have that:*

$$\hat{\tau}_{GRDD}(Z, \eta) \xrightarrow{p} \tau^*_{GRDD}(Z, \eta) \tag{35}$$

for all η and $Z \neq Z^*$. Further, if we assume that $\omega_k(Z, Z_i, k') = 0$ if $k \neq k'$, $\lambda_k = \lambda$ for all k and that $\lambda^{-1}\sqrt{\lambda^{-1}n} \to 0$, we get that:

$$\sqrt{\lambda^{-1}n} \cdot \Big(\hat{\tau}_{GRDD}(Z,\eta) - \tau^*_{GRDD}(Z,\eta)\Big) \to N(0, V) \tag{36}$$

where the variance is equal to:

$$\begin{aligned} V = {} & \mathbb{V}\big(\hat{\tau}_{obs}(Z)\big) + \mathbb{V}\big(\hat{\nu}(Z)\big) \cdot \big(\beta^*_1 - \beta^*_0\big)^2 \\ & + \mathbb{V}\big(\hat{\Delta}Y_1\big) \cdot \left(\frac{\nu(Z) - 2\eta}{\Delta p}\right)^2 + \\ & + \mathbb{V}\big(\hat{\Delta}Y_0\big) \cdot \left(\frac{\nu(Z) - 2\eta + 1}{\Delta p}\right)^2 \\ & + \mathbb{V}\big(\hat{\Delta}p\big) \cdot \left(\frac{\big(\nu(Z) - 2\eta\big) \cdot \big(\beta^*_1 - \beta^*_0\big) - \beta^*_0}{\Delta p}\right)^2 \end{aligned} \tag{37}$$

and where:

$$\hat{\Delta}Y_1 = \sum_{\forall i} \omega_1(\Delta Z^*, Z_i, T_i) \cdot Y_i \tag{38}$$

$$\hat{\Delta}Y_0 = \sum_{\forall i} \omega_0(\Delta Z^*, Z_i, T_i) \cdot Y_i \tag{39}$$

$$\hat{\Delta}p = \sum_{\forall i} \omega_\nu(\Delta Z^*, Z_i) \cdot T_i \tag{40}$$

Proof. From Assumption 6, we get that:

$$\sqrt{\lambda^{-1}n} \cdot \left(\begin{bmatrix} \hat{\tau}_{obs}(Z) - \tau^*_{obs}(Z) - b_1 \\ \hat{\nu}(Z) - \nu(Z) - b_2 \\ \hat{\Delta}Y_0 - \Delta Y_0 - b_3 \\ \hat{\Delta}Y_1 - \Delta Y_1 - b_4 \\ \hat{\Delta}p - \Delta p - b_5 \end{bmatrix} \right) \to N(0, \Sigma) \tag{66}$$

where the b_k terms are bias terms with $b_k \in [0, \infty)$.

Next, from Assumption 5, we get that the conditional moments estimated at Z are independent from the conditional moments estimated at Z^*, and so we can infer

that:

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \tag{67}$$

where Σ_1 is a two-by-two symmetric positive semi-definite matrix and Σ_2 is a four-by-four symmetric positive semi-definite matrix.

We next consider the covariance between the estimates of $\hat{\Delta}Y_1$ and $\hat{\Delta}Y_0$, which (because of the assumption that the observations are independent) we can write as:

$$Cov(\hat{\Delta}Y_1, \hat{\Delta}Y_0) = \sum_{\forall i} \Big(\omega_1(\Delta Z^*, Z_i, T_i) \cdot \omega_0(\Delta Z^*, Z_i, T_i) \Big) \cdot Var(Y_i | Z_i) \tag{68}$$

which equals zero, since $\omega_1(\Delta Z^*, Z_i, T_i) \cdot \omega_0(\Delta Z^*, Z_i, T_i) = 0$ regardless of T_i.

Finally, we get that the covariance between $\hat{\Delta}Y_1$ and $\hat{\Delta}p$ is zero from the fact that we can write $Y_i = y_1^*(\nu(Z_i), Z_i) \cdot T_i + y_0^*(\nu(Z_i), Z_i) \cdot (1 - T_i) + e_i$ with $\mathbb{E}[e_i | T_i, Z_i] = 0$ and that we can ignore the biases. To see how, note that we can write that:

$$\begin{aligned} \hat{\Delta}Y_1 &= \sum_{\forall i} \omega_1(\Delta Z^*, Z_i, T_i) \cdot y_1^*(\nu(Z_i), Z_i) + \sum_{\forall i} \omega_1(\Delta Z^*, Z_i, T_i) \cdot e_i \\ &= \Delta Y_1 + \Big(\sum_{\forall i} \omega_1(\Delta Z^*, Z_i, T_i) \cdot y_1^*(\nu(Z_i), Z_i) - \Delta Y_1 \Big) + \sum_{\forall i} \omega_1(\Delta Z^*, Z_i, T_i) \cdot e_i \\ &= \Delta Y_1 + b_{\hat{\Delta}Y_1} + e_{\hat{\Delta}Y_1} \end{aligned} \tag{69}$$

with $b_{\hat{\Delta}Y_1}$ corresponding to the bias term and $e_{\hat{\Delta}Y_1}$ to the error term with, importantly, $\mathbb{E}[e_{\hat{\Delta}Y_1} | Z^n, T^n] = 0$. We can similarly write $\hat{\Delta}p = \Delta_p + b_{\hat{\Delta}p} + e_{\hat{\Delta}p}$ for some bias term $b_{\hat{\Delta}p}$ and error term with $\mathbb{E}[e_{\hat{\Delta}p} | Z^n] = 0$. We therefore get that:

$$\begin{aligned} Cov(\hat{\Delta}Y_1, \hat{\Delta}p) &= \mathbb{E}[\hat{\Delta}Y_1 \hat{\Delta}p] - \mathbb{E}[\hat{\Delta}Y_1]\mathbb{E}[\hat{\Delta}p] \\ &= \mathbb{E}\big[\big(\Delta Y_1 + b_{\hat{\Delta}Y_1} + e_{\hat{\Delta}Y_1} \big) \cdot \big(\Delta_p + b_{\hat{\Delta}p} + e_{\hat{\Delta}p} \big) \big] - \mathbb{E}\big[\big(\Delta Y_1 + b_{\hat{\Delta}Y_1} \big) \cdot \big(\Delta_p + b_{\hat{\Delta}p} \big) \big] \\ &= \mathbb{E}\big[e_{\hat{\Delta}Y_1} \cdot e_{\hat{\Delta}p} \big] + \mathbb{E}\big[b_{\hat{\Delta}Y_1} \cdot e_{\hat{\Delta}p} \big] \end{aligned} \tag{70}$$

where the last equation comes form the fact that ΔY_1, Δp, and $b_{\hat{\Delta}p}$ are all functions of Z^n (which we hold fixed for the expectation) and not T^n or Y^n (which the expectations are taken over). We can further get that: $\mathbb{E}[e_{\hat{\Delta}Y_1} \cdot e_{\hat{\Delta}p}] = \mathbb{E}[\mathbb{E}[e_{\hat{\Delta}Y_1} | T^n] \cdot e_{\hat{\Delta}p}] = 0$, since $\mathbb{E}[e_{\hat{\Delta}Y_1} | T^n] = 0$ for all T^n. Finally, from the assumption that $\lambda^{-1}\sqrt{\lambda^{-1}n} \to 0$ we can

get that $b_{\hat{\Delta}Y_1} = 0$ and so $Cov(\hat{\Delta}Y_1, \hat{\Delta}p) = 0$.

We can use a similar approach to show the rest of the covariances are likewise equal to zero and the stated result then follows via the delta-method.

□

Proposition 10. *Let $\hat{\tau}_{GRDD}$ be the estimate generated by the Global Regression Discontinuity Design, as defined in Definition 2, and τ^*_{GRDD} be the function defined in Definition 1. Next, suppose that $\int_0^1 \omega(\eta, Z)d\eta = \omega_z(Z)$ for some continuous $\omega_z(Z) \in [0, \infty)$ with $\int_{\mathbf{Z}} \omega_z(Z) = 1$.*

Then under the assumptions in Proposition 9, we get that:

$$\sqrt{\lambda^{-1}n}\cdot\left(\int_{\boldsymbol{Z}}\int_0^1 \hat{\tau}_{GRDD}(\eta, Z)\omega(\eta, Z)d\eta dZ - \int_{\boldsymbol{Z}}\int_0^1 \tau^*_{GRDD}(\eta, Z)\omega(\eta, Z)d\eta dZ\right) \to N(0, V) \tag{41}$$

where the variance is equal to:

$$\begin{aligned} V = \mathbb{V}\big(\hat{\Delta}Y_1\big)\cdot\left(\frac{\xi}{\Delta p}\right)^2 + \mathbb{V}\big(\hat{\Delta}Y_0\big)\cdot\left(\frac{1+\xi}{\Delta p}\right)^2 \\ + \mathbb{V}\big(\hat{\Delta}p\big)\cdot\left(\frac{\xi\cdot\big(\beta^*_1 - \beta^*_0\big) - \beta^*_0}{\Delta p}\right)^2 \end{aligned} \tag{42}$$

and where $\xi = \int_{\boldsymbol{Z}}\int_0^1 \big(\nu(Z) - 2\eta\big)\omega(\eta, Z)d\eta dZ$.

Proof. From the formulation of the estimator, we get that:

$$\begin{aligned} \int_{\mathbf{Z}}\int_0^1 \hat{\tau}_{GRDD}(\eta, Z)\omega(\eta, Z)d\eta dZ = \int_{\mathbf{Z}} \hat{\tau}_{obs}(Z)\omega_z(Z)dZ \\ - \big(\hat{\beta}_1 - \hat{\beta}_0\big)\cdot\int_{\mathbf{Z}}\big(\hat{\nu}(Z) - 1\big)\omega_z(Z)dZ \\ - \hat{\beta}_0 \end{aligned} \tag{71}$$

From the assumption that C_0 and C_1 are twice-continuously differentiable, we can approximate the integrals $\int_{\mathbf{Z}} \hat{\tau}_{obs}(Z)\omega_z(Z)dZ$ and $\int_{\mathbf{Z}}\big(\hat{\nu}(Z) - 1\big)\omega_z(Z)dZ$ via Riemann sums. Further, from Assumption 5, the $\hat{\tau}(Z)$ and $\hat{\nu}(Z)$ estimates in these sums can be thought of as independent random variables with finite variance. Thus, since the weight on any one goes to zero as the number of terms increases, the variance of the sum vanishes. Thus, the only uncertainty in the estimate stems from uncertainty in $\hat{\beta}_0$ and $\hat{\beta}_1$.

□

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. 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. In particular, we focus here on how the associated R package handles these extensions and it to future papers to study whether there are better ways to extend the model.

B.A Covariates

As mentioned in Section III, the assumption required for τ^*_{GRDD} to be equal to the true MTE function will 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 and one can do so by simply saturating the model with covariate interactions. In practice, however, estimating the conditional moments non-parametrically, as we do in Section IV, quickly gets challenging as the number of covariates increases and so implementing the method requires 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 very similar to the discussion in Brinch et al. (2017), for example.

We view this, and especially 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. The most straightforward approach, however, is to extend the restriction on the conditional moments to be that:

$$\hat{y}_k(\eta, Z, X) = \gamma_k(Z, X) + \beta_k \eta \tag{72}$$

for $k \in \{0, 1\}$. We can then extend the formulation that the estimator is equal to a

bias-adjusted observational study to be that:

$$\tau_{GRDD}(\eta, Z, X) = \tau_{obs}(Z, X) - \big(\nu(Z, X) - 2\eta\big) \cdot \big(\beta_1 - \beta_0\big) - \beta_0 \tag{73}$$

Here, the formulation that the estimator is equal to a bias-adjusted observational study is quite useful if one wants to adjust for multiple covariates, since there are a number of studies and statistical packages that discuss how best to flexibly estimate conditional average treatment effects in an observational study context. Thus, one can estimate $\hat{\tau}_{obs}(Z, X)$ and $\hat{\nu}(Z, X) = \mathbb{E}[T_i | Z_i = Z, X_i = X]$ using standard parametric or non-parametric approaches.

Note, however, that even under the assumption that $\gamma_k(Z, X)$ is a continuous function and that the pdf $f(X|Z)$ is continuous around the discontinuity, it does not follow that $\beta_k = \frac{lim_{Z \uparrow Z^*} \mathbb{E}[Y_i | Z_i = Z, T_i = k] - lim_{Z \downarrow Z^*} \mathbb{E}[Y_i | Z_i = Z, T_i = k]}{lim_{Z \uparrow Z^*} \mathbb{E}[T_i | Z_i = Z] - lim_{Z \downarrow Z^*} \mathbb{E}[T_i | Z_i = Z]}$ for $k = \{0, 1\}$. This is because even if the pdf $f(X|Z)$ is continuous around the discontinuity, the conditional distribution $f(X|Z, T = k)$ will not be if the variables X affect the probability of treatment. Despite this complication there are multiple ways to estimate β_k, such as a covariate-adjusted regression discontinuity. Combining these estimates with the estimates of $\hat{\tau}_{obs}(Z, X)$ and $\hat{\nu}(Z, X)$ allows for the estimation of a covariate-adjusted Global RDD.

B.B Multiple Discontinuities

It is often the case that 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 we specify the model as having a single discontinuity 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 $\hat{y}_k$ that is additively seperable and linear in η such that $\hat{y}_k(\nu(Z), Z) = \mathbb{E}[Y_i | T_i = k, Z_i = Z]$ for all $Z \neq Z^*$ and $k \in \{0, 1\}$. With multiple discontinuities, in contrast, 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 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 restriction that $\hat{y}_k(\eta, Z)$ are additively separable and linear in η even in the presence of multiple discontinuities. While this complicates some of the motivation inherent in the Bayesian model, combining the discontinuities may lead to a more precise estimation of the "average slope" and we believe this will likely preferable in most contexts. It is this approach that we implement in the R package.

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.C 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.[28] 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 Imbens (2020) for more details on the test.[29] 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 (11) and (12).

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 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

[28] Many thanks to the participants at the 2023 AEFP for their thoughtful comments on a (very!) early version of this paper which initially motivated this section and to an anonymous referee who helped convince us to include it in the paper.

[29] 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).

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.

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 approaches that combine 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 how a MTE specification of the RD model can provide guidance on how to best combine noisy and local fuzzy RD estimates with biased but precise propensity score estimates.

C Appendix Figures and Tables

Table 2: Average Effect with Different Treatment Assignments

	Optimal Assignment	Realized Assignment (ATT)	Local Effect (LATE)	Random Assignment (ATE)	Program Expansion (ATC)
Grade 4	1.15	1.14	0.88	0.24	0.13
	(0.09)	(0.09)	(0.09)	(0.26)	(0.29)
Grade 5	0.90	0.87	0.72	0.52	0.48
	(0.10)	(0.10)	(0.09)	(0.23)	(0.25)
Grade 6	0.69	0.68	0.57	0.21	0.15
	(0.10)	(0.09)	(0.09)	(0.22)	(0.25)
Grade 7	0.52	0.51	0.41	0.14	0.09
	(0.08)	(0.08)	(0.07)	(0.20)	(0.22)
Grade 8	0.59	0.57	0.48	0.35	0.33
	(0.11)	(0.07)	(0.07)	(0.26)	(0.29)

Note: Standard errors are shown in parentheses and are clustered at the school level. 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).

Table 3: Estimates of How the LATE Differs from Other Estimands

	ATE vs LATE	LATE vs ATC
Grade 4	0.16	0.90
	(0.03)	(0.32)
Grade 5	0.09	0.41
	(0.04)	(0.24)
Grade 6	-0.01	0.11
	(0.03)	(0.22)
Grade 7	-0.00	0.12
	(0.03)	(0.21)
Grade 8	-0.02	-0.09
	(0.03)	(0.25)

Note: Standard errors are shown in parentheses and are clustered at the school level. ATE vs LATE is the estimated difference between the average treatment on the treated (ATT) and the average effect of the program on compliers at the treatment threshold (LATE). LATE vs ATC is the difference between the LATE and the average treatment on the controls (ATC).

Table 4: Global RDD vs. RD Robust

	Global RDD	RD Robust
Grade 4	0.87	1.09
	(0.09)	(0.30)
Grade 5	0.67	0.78
	(0.10)	(0.41)
Grade 6	0.48	0.76
	(0.09)	(0.45)
Grade 7	0.36	0.25
	(0.08)	(0.24)
Grade 8	0.35	0.84
	(0.08)	(0.44)

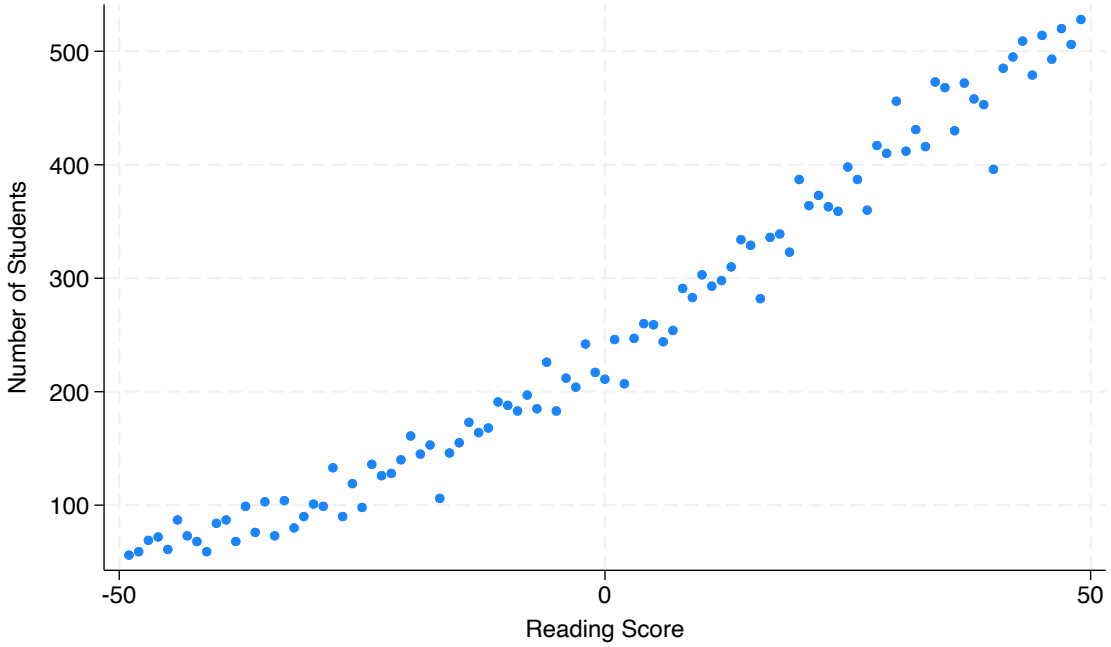
Note: Standard errors are shown in parentheses and are clustered at the school level. Both estimates are of the LATE, i.e., the average effect of the program on compliers at the treatment threshold. The Global RDD is estimated using the approach outlined in the paper, with RD Robust using the approach of .

Table 5: Estimated Effects When Including Students with a Documented Disability

	Optimal Assignment	Realized Assignment (ATT)	Local Effect (LATE)	Random Assignment (ATE)	Program Expansion (ATC)
Grade 4	1.29	1.14	0.90	0.90	0.87
	(0.10)	(0.09)	(0.08)	(0.10)	(0.11)
Grade 5	0.90	0.82	0.68	0.59	0.56
	(0.09)	(0.08)	(0.08)	(0.08)	(0.08)
Grade 6	0.62	0.61	0.52	0.33	0.30
	(0.08)	(0.08)	(0.07)	(0.13)	(0.14)
Grade 7	0.48	0.45	0.36	0.28	0.26
	(0.07)	(0.06)	(0.06)	(0.07)	(0.07)
Grade 8	0.61	0.56	0.48	0.42	0.40
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)

Note: Standard errors are shown in parentheses and are clustered at the school level. 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).

Figure 6: Density Around the Distribution



Note: The figure plots how many students scored each possible value of the third-grade reading test around the proficiency threshold.

Table 6: Additional Monte Carlo Results

(a) Estimates of LATE

	Squared Bias		Variance		Mean-Squared Error	
Sample Size	rdrobust	Global RD	rdrobust	Global RD	rdrobust	Global RD
1,000	9.53	0.10	659.8	1.10	669.4	1.20
2,500	1.53	0.03	126.6	0.36	128.2	0.39
5,000	0.16	0.01	9.77	0.17	9.94	0.19
10,000	0.11	0.006	0.75	0.09	0.76	0.10

(b) Estimates of ATE

	Squared Bias		Variance		Mean-Squared Error	
Sample Size	PSWeight	Global RD	PSWeight	Global RD	PSWeight	Global RD
1,000	0.70	0.27	0.007	5.88	0.71	6.16
2,500	0.70	0.12	0.003	0.67	0.71	0.80
5,000	0.70	0.10	0.001	0.25	0.71	0.35
10,000	0.70	0.09	0.0007	0.13	0.70	0.13

Note: This table shows the results of the Monte Carlo simulation described in Section IV.C. In it, we generate 50 true conditional moments and for each of these, simulate the rest of the data generating process 100 times and each time estimate the treatment effects. We can then calculate the average squared-bias, variance, and mean-squared error. "rdrobust" is a traditional RDD estimate that uses the rdrobust package to calculate the bandwidth (Calonico et al., 2015a). "PSWeight" is a propensity score weighting approach that estimates the true propensity scores using a simple logit model. "Global RD" is the approach outlined in this paper. The simulation results with other alternatives are shown in Appendix Table 1.